

Comparing two treatments in terms of the likelihood ratio order

Martin, N.¹, Mata, R.² and Pardo, L.²

¹Department of Statistics, Carlos III University of Madrid, 28903 Getafe (Madrid), Spain

²Department of Statistics and O.R., Complutense University of Madrid, 28040 Madrid, Spain

February 27, 2014

Abstract

In this paper new families of test statistics are introduced and studied for the problem of comparing two treatments in terms of the likelihood ratio order. The considered families are based on phi-divergence measures and arise as natural extensions of the classical likelihood ratio test and Pearson test statistics. It is proven that their asymptotic distribution is a common chi-bar random variable. An illustrative example is presented and the performance of these statistics is analysed through a simulation study. Through a simulation study it is shown that, for most of the proposed scenarios adjusted to be small or moderate, some members of this new family of test-statistic display clearly better performance with respect to the power in comparison to the classical likelihood ratio and the Pearson's chi-square test while the exact size remains closed to the nominal size.

Keywords and phrases: Divergence measure, Kullback divergence measure, Inequality constraints, Likelihood ratio order, Loglinear models.

1 Introduction

In Table 1 the results of an experiment to compare two treatments for ulcer is shown. This article proposes new families of test-statistics when we are interested in studying the possibility that the treatment is better than the control.

	Larger	$< \frac{1}{3}$	Healed	$> \frac{2}{3}$	Healed	Healed
Treatment 1 (Control)	12		10		4	6
Treatment 2 (Treatment)	5		8		8	11

Table 1: Change in size of ulcer crater.

Let Y denote the ordinal response variable and X denote an ordinal explanatory variable with two categories. The variable Y takes the values 1, 2, 3 and 4, which represent different levels of healing, from less to much capacity to heal the ulcer. The variable X takes the values 1 and 2 according as the treatment group, 1 is control and 2 is the treatment group by itself. We shall initially focus on making statistical inference on the theoretical probabilities displayed in Table 2.

There are several ways of formulating the statement “the treatment is better than the control”. Initially, we shall consider that Treatment 2 is at least as good as Treatment 1 if the ratio $\frac{\Pr(Y=j|X=2)}{\Pr(Y=j|X=1)}$ increases as the response category, j , increases, i.e.

$$\frac{\Pr(Y=j|X=2)}{\Pr(Y=j|X=1)} \leq \frac{\Pr(Y=j+1|X=2)}{\Pr(Y=j+1|X=1)} \quad \text{for every } j, \quad (1)$$

	Larger	$< \frac{1}{3}$ Healed	$> \frac{2}{3}$ Healed	Healed
Treatment 1 (Control)	$\Pr(Y = 1 X = 1)$	$\Pr(Y = 2 X = 1)$	$\Pr(Y = 3 X = 1)$	$\Pr(Y = 4 X = 1)$
Treatment 2 (Treatment)	$\Pr(Y = 1 X = 2)$	$\Pr(Y = 2 X = 2)$	$\Pr(Y = 3 X = 2)$	$\Pr(Y = 4 X = 2)$

Table 2: Theoretical conditional probabilities.

and Treatment 2 is better than the Treatment 1 if (1) holds with at least one strict inequality.

If we assume that Treatment 2 is at least as good as Treatment 1, i.e., (1) holds, is there any evidence to support the claim that treatment 2 is better? In such a case null and alternative hypotheses may be

$$H_0 : \frac{\Pr(Y=j|X=2)}{\Pr(Y=j|X=1)} = \frac{\Pr(Y=j+1|X=2)}{\Pr(Y=j+1|X=1)} \quad \text{for every } j, \quad (2a)$$

$$H_1 : \frac{\Pr(Y=j|X=2)}{\Pr(Y=j|X=1)} \leq \frac{\Pr(Y=j+1|X=2)}{\Pr(Y=j+1|X=1)} \quad \text{for every } j \quad \text{and} \quad \frac{\Pr(Y=j|X=2)}{\Pr(Y=j|X=1)} < \frac{\Pr(Y=j+1|X=2)}{\Pr(Y=j+1|X=1)} \quad \text{for at least one } j. \quad (2b)$$

The null hypothesis means that both treatments are equally effective, while the alternative hypothesis means that Treatment 2 is more effective than Treatment 1. Note that if we multiply on the left and right hand side of (2a) and (2b) by $\left(\frac{\Pr(Y=j|X=2)}{\Pr(Y=j|X=1)}\right)^{-1}$ we obtain

$$\begin{aligned} H_0 : \vartheta_j &= 1 \quad \text{for every } j \in \{1, \dots, J-1\}, \\ H_1 : \vartheta_j &\geq 1 \quad \text{for every } j \in \{1, \dots, J-1\} \quad \text{and} \quad \vartheta_j > 1 \quad \text{for at least one } j \in \{1, \dots, J-1\}, \end{aligned}$$

where J is the number of ordered categories for response variable Y ,

$$\vartheta_j = \frac{\pi_{1j}\pi_{2,j+1}}{\pi_{2j}\pi_{1,j+1}}, \quad \forall j \in \{1, \dots, J-1\}, \quad (4)$$

are “local odds ratios” associated with response category j , and

$$\pi_{ij} = \Pr(Y = j|X = i). \quad (5)$$

The non-parametric statistical inference associated with the likelihood ratio ordering for two multinomial samples was introduced for the first time in Dykstra et al. (1995) using the likelihood ratio test-statistic. In the literature related to different types of orderings, in general there is not very clear what is the most appropriate ordering to compare two treatments according to a categorized ordinal variable. In the case of having two independent multinomial samples, the likelihood ratio ordering is the most restricted ordering type; for example, if the likelihood ratio ordering holds, then the simple stochastic ordering also holds. Dardanoni and Forcina (1998) proposed a new method for making statistical inference associated with different types of orderings. For unifying and comparing different types of orderings, they reparametrize the initial model. Different ordering types can be considered to be nested models and the likelihood ratio ordering is the most parsimonious one. The advantage of nested models is that the most restricted models tend to be more powerful for the alternatives that belong to the most restricted alternatives. In this setting, our proposal in this paper is to introduce new test-statistics that provide substantially better power for testing (2a) against (2b).

The structure of the paper is as follows. In Section 2, we have considered the likelihood ratio order associated with a non-parametric model, as in Dardanoni and Forcina, but the specification of the model through a saturated loglinear model is substantially different. Section 3 presents the phi-divergence test-statistics as extension of the likelihood ratio and chi-square test-statistics. The applied methodology in Section 4 for proving the asymptotic distribution of the phi-divergence test-statistics, based on loglinear modeling, has been developed by following a completely new and meaningful method even for the likelihood ratio test. A numerical example is given in Section 5. The aim of Section 6 is to study through simulation the behaviour of the phi-divergence test-statistics for small and moderate sample sizes. Finally, we present an Appendix in which we establish the part of the proofs of the results not shown in Section 4.

2 Loglinear modeling

We display the whole distribution of π_{ij} , given in (5), in a rectangular table having 2 rows for the categories of X and J columns for the categories of Y (for the initial example, Table 2) and we denote the $2 \times J$ matrix $\Pi = (\pi_1, \pi_2)^T$, with two rows of probability vectors, $\pi_i = (\pi_{i1}, \dots, \pi_{iJ})^T$, $i = 1, 2$. We consider two independent random samples $\mathbf{N}_i = (N_{i1}, \dots, N_{iJ})^T \sim \mathcal{M}(n_i, \pi_i)$, $i = 1, 2$, where sizes n_i are prefixed and $\pi_i > \mathbf{0}_J$, that is the probability distribution of r.v. $\mathbf{N} = (\mathbf{N}_1^T, \mathbf{N}_2^T)^T$ is product-multinomial. Let

$$p_{ij} = \Pr(X = i, Y = j), \quad (6)$$

be the joint probability distribution. Since $\Pr(X = i, Y = j) = \Pr(Y = j|X = i) \Pr(X = i)$, i.e. $p_{ij} = \pi_{ij} \frac{n_i}{n}$, $i = 1, 2$, where $n = n_1 + n_2$, we can express (4) also in terms of the joint probabilities

$$\vartheta_j = \frac{p_{1j} p_{2,j+1}}{p_{2j} p_{1,j+1}}, \quad \forall j \in \{1, \dots, J-1\}. \quad (7)$$

Let $\mathbf{P} = (\mathbf{p}_1, \mathbf{p}_2)^T$, with $\mathbf{p}_i = (p_{i1}, \dots, p_{iJ})^T$, $i = 1, 2$, be the $2 \times J$ probability matrix and

$$\mathbf{p} = \text{vec}(\mathbf{P}^T) = (\mathbf{p}_1^T, \mathbf{p}_2^T)^T \quad (8)$$

a probability vector obtained by stacking the columns of \mathbf{P}^T (i.e., the rows of matrix \mathbf{P}). Note that the components of \mathbf{P} are ordered in lexicographical order in \mathbf{p} . The likelihood function of \mathbf{N} is $\mathcal{L}(\mathbf{N}; \mathbf{p}) = k \prod_{j=1}^J p_{1j}^{N_{1j}} p_{2j}^{N_{2j}}$, where k is a constant which does not depend on \mathbf{p} and the kernel of the loglikelihood function

$$\ell(\mathbf{N}; \mathbf{p}) = \sum_{j=1}^J (N_{1j} \log p_{1j} + N_{2j} \log p_{2j}). \quad (9)$$

In matrix notation, we are interested in testing

$$H_0 : \boldsymbol{\vartheta} = \mathbf{1}_{J-1} \text{ versus } H_1 : \boldsymbol{\vartheta} \not\geq \mathbf{1}_{J-1}, \quad (10)$$

where $\mathbf{1}_a$ is the a -vector of 1-s, $\boldsymbol{\vartheta} = (\vartheta_1, \dots, \vartheta_{J-1})^T$. Note that (10) involves $J-1$ non-linear constraints on \mathbf{p} , defined by (8). In this article the hypothesis testing problem is formulated making a reparametrization of \mathbf{p} using the saturated loglinear model, so that some linear restrictions are considered with respect to the new parameters. This fact is important and interesting.

Focussed on \mathbf{p} , the saturated loglinear model with canonical parametrization is defined by

$$\log p_{ij} = u + u_{1(i)} + \theta_{2(j)} + \theta_{12(ij)}, \quad (11)$$

with the identifiability restrictions

$$u_{1(2)} = 0, \quad \theta_{2(J)} = 0, \quad \theta_{12(1J)} = 0, \quad \theta_{12(2j)} = 0, \quad j = 1, \dots, J. \quad (12)$$

It is important to clarify that we have used the identifiability constraints (12) in order to make easier the calculations and this model formulation for making statistical inference with inequality restrictions with local odds-ratios has been given in this paper for the first time. Similar conditions have been used for instance in Lang (1996, examples of Section 7) and Silvapulle and Sen (2005, exercise 6.25 in page 345). Let $\boldsymbol{\theta}_{12} = (\theta_{12(11)}, \dots, \theta_{12(1,J-1)})^T$, $\boldsymbol{\theta}_2 = (\theta_{2(1)}, \dots, \theta_{2(J-1)})^T$ denote subvectors of the unknown parameters $\boldsymbol{\theta} = (\boldsymbol{\theta}_2^T, \boldsymbol{\theta}_{12}^T)^T$. The components of $\mathbf{u} = (u, u_{1(1)})^T$ are redundant parameters since the term u can be expressed in function of $\boldsymbol{\theta}$ using the fact that $\sum_{j=1}^J p_{2j} = \frac{n_2}{n}$, i.e.

$$u = u(\boldsymbol{\theta}) = \log n_2 - \log n - \log \left(1 + \sum_{j=1}^{J-1} \exp\{\theta_{2(j)}\} \right), \quad (13)$$

and $u_{1(1)}$ taking into account that $\sum_{j=1}^J p_{1j} = \frac{n_1}{n}$, i.e.

$$u_{1(1)} = u_{1(1)}(\boldsymbol{\theta}) = \log \frac{n_1}{n_2} + \log \frac{1 + \sum_{j=1}^{J-1} \exp\{\theta_{2(j)}\}}{1 + \sum_{j=1}^{J-1} \exp\{\theta_{2(j)} + \theta_{12(1j)}\}}. \quad (14)$$

In matrix notation (11) is given by

$$\log \mathbf{p}(\boldsymbol{\theta}) = \mathbf{W}_0 \mathbf{u} + \mathbf{W} \boldsymbol{\theta}, \quad (15)$$

where $\mathbf{p}(\boldsymbol{\theta})$ is \mathbf{p} such that the components are defined by (11),

$$\mathbf{W}_0 = \begin{pmatrix} 1 & 1 \\ 1 & 0 \end{pmatrix} \otimes \mathbf{1}_J$$

is a $2J \times 2$ matrix with $\mathbf{1}_a$ being the a -vector of ones, $\mathbf{0}_a$ the a -vector of zeros, \otimes the Kronecker product; \mathbf{W} the full rank design matrix of size $2J \times 2(J-1)$, such that

$$\mathbf{W} = \begin{pmatrix} 1 & 1 \\ 1 & 0 \end{pmatrix} \otimes \begin{pmatrix} \mathbf{I}_{J-1} \\ \mathbf{0}_{J-1}^T \end{pmatrix}, \quad (16)$$

with \mathbf{I}_a being the identity matrix of order a , $\mathbf{0}_{a \times b}$ the matrix of size $a \times b$ with zeros. The condition (1) can be expressed by the linear constraint

$$\theta_{12(1j)} - \theta_{12(2j)} - \theta_{12(1,j+1)} + \theta_{12(2,j+1)} \geq 0, \quad \forall j \in \{1, \dots, J-1\}, \quad (17)$$

since

$$\log \vartheta_j = \log p_{1j} - \log p_{2j} - \log p_{1,j+1} + \log p_{2,j+1} = \theta_{12(1j)} - \theta_{12(2j)} - \theta_{12(1,j+1)} + \theta_{12(2,j+1)}.$$

Condition (17) in matrix notation is given by $\mathbf{R}\boldsymbol{\theta} \geq \mathbf{0}_{J-1}$, with $\mathbf{R} = \mathbf{e}_2^T \otimes \mathbf{G}_{J-1} = (\mathbf{0}_{(J-1) \times (J-1)}, \mathbf{G}_{J-1})$, \mathbf{e}_a is the a -th unit vector and \mathbf{G}_h is a $h \times h$ matrix with 1-s in the main diagonal and -1-s in the upper superdiagonal. Observe that the restrictions can be expressed also as $\mathbf{G}_{J-1}\boldsymbol{\theta}_{12} \geq \mathbf{0}_{J-1}$, and $\theta_{1(1)}$ are $\boldsymbol{\theta}_2$ are nuisance parameters because they do not take part actively in the restrictions.

The kernel of the likelihood function with the new parametrization is obtained replacing \mathbf{p} by $\mathbf{p}(\boldsymbol{\theta})$ in (9), i.e.

$$\ell(\mathbf{N}; \boldsymbol{\theta}) = \mathbf{N}^T \log \mathbf{p}(\boldsymbol{\theta}) = \mathbf{N}^T (\mathbf{W}_0 \mathbf{u} + \mathbf{W} \boldsymbol{\theta}) = n\mathbf{u}(\boldsymbol{\theta}) + n_1 u_{1(1)}(\boldsymbol{\theta}) + \mathbf{N}^T \mathbf{W} \boldsymbol{\theta}.$$

Hypotheses (10) can be now formulated as

$$H_0 : \mathbf{R}\boldsymbol{\theta} = \mathbf{0}_{J-1} \text{ versus } H_1 : \mathbf{R}\boldsymbol{\theta} \geq \mathbf{0}_{J-1} \text{ and } \mathbf{R}\boldsymbol{\theta} \neq \mathbf{0}_{J-1}. \quad (18)$$

Under H_0 , the parameter space is $\Theta_0 = \{\boldsymbol{\theta} \in \mathbb{R}^{J+1} : \mathbf{R}\boldsymbol{\theta} = \mathbf{0}_{J-1}\}$ and the maximum likelihood estimator (MLE) of $\boldsymbol{\theta}$ in Θ_0 is $\hat{\boldsymbol{\theta}} = \arg \max_{\boldsymbol{\theta} \in \Theta_0} \ell(\mathbf{N}; \boldsymbol{\theta})$. The overall parameter space is $\Theta = \{\boldsymbol{\theta} \in \mathbb{R}^{J+1} : \mathbf{R}\boldsymbol{\theta} \geq \mathbf{0}_{J-1}\}$ and the MLE of $\boldsymbol{\theta}$ in Θ is $\tilde{\boldsymbol{\theta}} = \arg \max_{\boldsymbol{\theta} \in \Theta} \ell(\mathbf{N}; \boldsymbol{\theta})$. It is worthwhile to mention that the probability vectors for both parametric spaces, $\mathbf{p}(\hat{\boldsymbol{\theta}})$ and $\mathbf{p}(\tilde{\boldsymbol{\theta}})$ can be obtained by following the invariance property of the MLEs first estimating $\boldsymbol{\theta}$ and later plugging it into $\mathbf{p}(\boldsymbol{\theta})$, however $\mathbf{p}(\hat{\boldsymbol{\theta}})$ has an explicit expression,

$$p_{ij}(\hat{\boldsymbol{\theta}}) = \frac{n_i(N_{1j} + N_{2j})}{n^2}, \quad (19)$$

where $n_i = \sum_{j=1}^J N_{ij}$ (see Christensen (1997), Section 2.3, for more details).

3 Phi-divergence test-statistics

The likelihood ratio statistic for testing (10), equivalent to one given by Dykstra et al. (1995) but adapted for loglinear modeling, is

$$G^2 = 2(\ell(\mathbf{N}; \tilde{\boldsymbol{\theta}}) - \ell(\mathbf{N}; \hat{\boldsymbol{\theta}})) = 2n \sum_{i=1}^2 \sum_{j=1}^J \bar{p}_{ij} \log \frac{p_{ij}(\tilde{\boldsymbol{\theta}})}{p_{ij}(\hat{\boldsymbol{\theta}})}, \quad (20)$$

where $\bar{p}_{ij} = N_{ij}/n$, $i = 1, 2$, $j = 1, \dots, J$. Taking into account the identifiability constraints (12) and $\hat{u} = u(\hat{\boldsymbol{\theta}})$, $\tilde{u} = u(\tilde{\boldsymbol{\theta}})$, $\hat{u}_{1(1)} = u_{1(1)}(\hat{\boldsymbol{\theta}})$, $\tilde{u}_{1(1)} = u_{1(1)}(\tilde{\boldsymbol{\theta}})$ (see formulas (13)-(14)), (20) can also be expressed as

$$G^2 = 2n(\tilde{u} - \hat{u}) + 2n_1(\tilde{u}_{1(1)} - \hat{u}_{1(1)}) + 2\mathbf{N}^T \mathbf{W}(\tilde{\boldsymbol{\theta}} - \hat{\boldsymbol{\theta}}).$$

The chi-square statistic for testing (10) is

$$X^2 = n \sum_{i=1}^2 \sum_{j=1}^J \frac{(p_{ij}(\hat{\boldsymbol{\theta}}) - p_{ij}(\tilde{\boldsymbol{\theta}}))^2}{p_{ij}(\hat{\boldsymbol{\theta}})}. \quad (21)$$

The Kullback-Leibler divergence measure between two $2J$ -dimensional probability vectors \mathbf{p} and \mathbf{q} is defined as

$$d_{Kull}(\mathbf{p}, \mathbf{q}) = \sum_{i=1}^2 \sum_{j=1}^J p_{ij} \log \frac{p_{ij}}{q_{ij}}$$

and the Pearson divergence measure

$$d_{Pearson}(\mathbf{p}, \mathbf{q}) = \frac{1}{2} \sum_{i=1}^2 \sum_{j=1}^J \frac{(p_{ij} - q_{ij})^2}{q_{ij}}.$$

It is not difficult to check that

$$G^2 = 2n(d_{Kull}(\bar{\mathbf{p}}, \mathbf{p}(\hat{\boldsymbol{\theta}})) - d_{Kull}(\bar{\mathbf{p}}, \mathbf{p}(\tilde{\boldsymbol{\theta}}))) \quad (22)$$

and

$$X^2 = 2nd_{Pearson}(\mathbf{p}(\tilde{\boldsymbol{\theta}}), \mathbf{p}(\hat{\boldsymbol{\theta}})), \quad (23)$$

being $\bar{\mathbf{p}} = \mathbf{N}/n = (\bar{p}_{11}, \dots, \bar{p}_{1J}, \bar{p}_{21}, \dots, \bar{p}_{2J})^T$ the vector of relative frequencies.

More general than the Kullback-Leibler divergence and Pearson divergence measures are ϕ -divergence measures, defined as

$$d_{\phi}(\mathbf{p}, \mathbf{q}) = \sum_{i=1}^2 \sum_{j=1}^J q_{ij} \phi \left(\frac{p_{ij}}{q_{ij}} \right),$$

where $\phi : \mathbb{R}_+ \rightarrow \mathbb{R}$ is a convex function such that

$$\phi(1) = \phi'(1) = 0, \phi''(1) > 0, 0\phi\left(\frac{0}{0}\right) = 0, 0\phi\left(\frac{p}{0}\right) = p \lim_{u \rightarrow \infty} \frac{\phi(u)}{u}, \text{ for } p \neq 0.$$

From a statistical point of view, the first asymptotic statistical results based on divergence measures in multinomial populations were obtained in Zografos et al. (1990). For more details about ϕ -divergence measures see Pardo (2006) and Cressie and Pardo (2002).

Apart from the likelihood ratio statistic (20) and the chi-square (21) statistic, we shall consider two new families of test-statistics based on ϕ -divergence measures. The first new family is obtained by replacing in (22) the Kullback divergence measure by a ϕ -divergence measure,

$$T_{\phi}(\bar{\mathbf{p}}, \mathbf{p}(\tilde{\boldsymbol{\theta}}), \mathbf{p}(\hat{\boldsymbol{\theta}})) = \frac{2n}{\phi''(1)} (d_{\phi}(\bar{\mathbf{p}}, \mathbf{p}(\hat{\boldsymbol{\theta}})) - d_{\phi}(\bar{\mathbf{p}}, \mathbf{p}(\tilde{\boldsymbol{\theta}}))). \quad (24)$$

The second new family is obtained by replacing in (23) the Pearson divergence measure by a ϕ -divergence measure,

$$S_\phi(\mathbf{p}(\tilde{\boldsymbol{\theta}}), \mathbf{p}(\hat{\boldsymbol{\theta}})) = \frac{2n}{\phi''(1)} d_\phi(\mathbf{p}(\tilde{\boldsymbol{\theta}}), \mathbf{p}(\hat{\boldsymbol{\theta}})). \quad (25)$$

If we consider $\phi(x) = x \log x - x + 1$ in (24), we get G^2 , and if we consider $\phi(x) = \frac{1}{2}(x-1)^2$ in (24), we get X^2 . Test-statistics based on ϕ -divergence measures have been used in the framework of loglinear models for some authors, see Cressie and Pardo (2000, 2002, 2003), Martín and Pardo (2006, 2008b, 2011).

4 Asymptotic results

As starting point, we shall establish the observed Fisher information matrix associated with $\boldsymbol{\theta}$, $\mathcal{I}_F^{(n_1, n_2)}(\boldsymbol{\theta})$, for a loglinear model with product-multinomial sampling as

$$\mathcal{I}_F^{(n_1, n_2)}(\boldsymbol{\theta}) = \frac{1}{n} \mathbf{W}^T \begin{pmatrix} n_1(\mathbf{D}_{\boldsymbol{\pi}_1(\boldsymbol{\theta})} - \boldsymbol{\pi}_1(\boldsymbol{\theta})\boldsymbol{\pi}_1^T(\boldsymbol{\theta})) & \mathbf{0}_{J \times J} \\ \mathbf{0}_{J \times J} & n_2(\mathbf{D}_{\boldsymbol{\pi}_2(\boldsymbol{\theta})} - \boldsymbol{\pi}_2(\boldsymbol{\theta})\boldsymbol{\pi}_2^T(\boldsymbol{\theta})) \end{pmatrix} \mathbf{W}, \quad (26)$$

where $\mathbf{D}_{\mathbf{a}}$ is the diagonal matrix of vector \mathbf{a} . To proof (26), we take into account that the overall observed Fisher information matrix for product multinomial sampling is the weighted observed Fisher information matrix associated with each multinomial sample, $\mathcal{I}_{F,i}^{(n_1, n_2)}(\boldsymbol{\theta})$, $i = 1, 2$, i.e.

$$\begin{aligned} \mathcal{I}_F^{(n_1, n_2)}(\boldsymbol{\theta}) &= \frac{n_1}{n} \mathcal{I}_{F,1}^{(n_1, n_2)}(\boldsymbol{\theta}) + \frac{n_2}{n} \mathcal{I}_{F,2}^{(n_1, n_2)}(\boldsymbol{\theta}), \\ \mathcal{I}_{F,i}^{(n_1, n_2)}(\boldsymbol{\theta}) &= \mathbf{W}_i^T (\mathbf{D}_{\boldsymbol{\pi}_i(\boldsymbol{\theta})} - \boldsymbol{\pi}_i(\boldsymbol{\theta})\boldsymbol{\pi}_i^T(\boldsymbol{\theta})) \mathbf{W}_i, \quad i = 1, 2, \end{aligned}$$

such that $\mathbf{W}^T = (\mathbf{W}_1^T, \mathbf{W}_2^T)$, $\log \mathbf{p}_1(\boldsymbol{\theta}) = u\mathbf{1}_J + u_{1(1)}\mathbf{1}_J + \mathbf{W}_1\boldsymbol{\theta}$ and $\log \mathbf{p}_2(\boldsymbol{\theta}) = u\mathbf{1}_J + \mathbf{W}_2\boldsymbol{\theta}$.

When $\boldsymbol{\theta} \in \Theta_0$, we shall denote $\boldsymbol{\theta}_0$ to be the true value of the unknown parameter under H_0 , and in such a case it holds $\boldsymbol{\pi}_1(\boldsymbol{\theta}_0) = \boldsymbol{\pi}_2(\boldsymbol{\theta}_0) = \boldsymbol{\pi}(\boldsymbol{\theta}_0) = (\pi_1(\boldsymbol{\theta}_0), \dots, \pi_J(\boldsymbol{\theta}_0))^T$, where $\boldsymbol{\pi}_i(\boldsymbol{\theta}_0)$ is defined as the probability vector with the terms given in (5) and related to the loglinear model through $\mathbf{p}_i(\boldsymbol{\theta}_0) = \frac{n_i}{n} \boldsymbol{\pi}_i(\boldsymbol{\theta}_0)$, $i = 1, 2$. Notice that $\boldsymbol{\pi}_i(\boldsymbol{\theta}_0)$ is fixed as $n_1, n_2 \rightarrow \infty$ and we shall assume that

$$\nu_i = \lim_{n_i \rightarrow \infty} \frac{n_i}{n}, \quad i = 1, 2,$$

is fixed but unknown, i.e. $\lim_{n_i \rightarrow \infty} \mathbf{p}_i(\boldsymbol{\theta}) = \nu_i \boldsymbol{\pi}_i(\boldsymbol{\theta}_0)$, $i = 1, 2$. We shall also denote

$$\boldsymbol{\pi}^*(\boldsymbol{\theta}_0) = (\pi_1(\boldsymbol{\theta}_0), \dots, \pi_{J-1}(\boldsymbol{\theta}_0))^T, \quad i = 1, 2,$$

the $(J-1)$ -dimensional vector obtained removing from $\boldsymbol{\pi}(\boldsymbol{\theta}_0)$ the last element. Focussing on the parameter structure $\boldsymbol{\theta} = (\boldsymbol{\theta}_{12}^T, \boldsymbol{\theta}_2^T)^T$, with $\boldsymbol{\theta}_{12} = (\theta_{12(11)}, \dots, \theta_{12(1, J-1)})^T$, $\boldsymbol{\theta}_2 = (\theta_{2(1)}, \dots, \theta_{2(J-1)})^T$ and the specific structure of \mathbf{W} , see (16), we shall establish asymptotically the specific shape of (26), a fundamental result for the posterior theorems.

Theorem 1 *The asymptotic Fisher information matrix of $\boldsymbol{\theta}$, $\mathcal{I}_F(\boldsymbol{\theta}) = \lim_{n_1, n_2 \rightarrow \infty} \mathcal{I}_F^{(n_1, n_2)}(\boldsymbol{\theta})$ when $\boldsymbol{\theta} \in \Theta_0$ is given by*

$$\mathcal{I}_F(\boldsymbol{\theta}_0) = \begin{pmatrix} \mathbf{D}_{\boldsymbol{\pi}^*(\boldsymbol{\theta}_0)} - \boldsymbol{\pi}^*(\boldsymbol{\theta}_0)\boldsymbol{\pi}^{*T}(\boldsymbol{\theta}_0) & \nu_1 (\mathbf{D}_{\boldsymbol{\pi}^*(\boldsymbol{\theta}_0)} - \boldsymbol{\pi}^*(\boldsymbol{\theta}_0)\boldsymbol{\pi}^{*T}(\boldsymbol{\theta}_0)) \\ \nu_1 (\mathbf{D}_{\boldsymbol{\pi}^*(\boldsymbol{\theta}_0)} - \boldsymbol{\pi}^*(\boldsymbol{\theta}_0)\boldsymbol{\pi}^{*T}(\boldsymbol{\theta}_0)) & \nu_1 (\mathbf{D}_{\boldsymbol{\pi}^*(\boldsymbol{\theta}_0)} - \boldsymbol{\pi}^*(\boldsymbol{\theta}_0)\boldsymbol{\pi}^{*T}(\boldsymbol{\theta}_0)) \end{pmatrix}. \quad (27)$$

Proof. Replacing $\boldsymbol{\theta}$ by $\boldsymbol{\theta}_0$ and the explicit expression of \mathbf{W} in the general expression of the finite sample size Fisher information matrix for two independent multinomial samples, (26), we obtain through the property of

the Kronecker product given in (1.22) of Harville (2008, page 341) that

$$\begin{aligned}
\mathcal{I}_F^{(n_1, n_2)}(\boldsymbol{\theta}_0) &= \left(\begin{pmatrix} 1 & 1 \\ 1 & 0 \end{pmatrix} \otimes \begin{pmatrix} \mathbf{I}_{J-1} \\ \mathbf{0}_{J-1}^T \end{pmatrix}^T \right) \left(\text{diag}\left\{ \frac{n_i}{n} \right\}_{i=1}^2 \otimes (\mathbf{D}_{\boldsymbol{\pi}(\boldsymbol{\theta}_0)} - \boldsymbol{\pi}(\boldsymbol{\theta}_0)\boldsymbol{\pi}^T(\boldsymbol{\theta}_0)) \right) \left(\begin{pmatrix} 1 & 1 \\ 1 & 0 \end{pmatrix} \otimes \begin{pmatrix} \mathbf{I}_{J-1} \\ \mathbf{0}_{J-1}^T \end{pmatrix} \right) \\
&= \left(\begin{pmatrix} 1 & 1 \\ 1 & 0 \end{pmatrix} \text{diag}\left\{ \frac{n_i}{n} \right\}_{i=1}^2 \begin{pmatrix} 1 & 1 \\ 1 & 0 \end{pmatrix} \right) \otimes \left(\begin{pmatrix} \mathbf{I}_{J-1} \\ \mathbf{0}_{J-1}^T \end{pmatrix}^T (\mathbf{D}_{\boldsymbol{\pi}(\boldsymbol{\theta}_0)} - \boldsymbol{\pi}(\boldsymbol{\theta}_0)\boldsymbol{\pi}^T(\boldsymbol{\theta}_0)) \begin{pmatrix} \mathbf{I}_{J-1} \\ \mathbf{0}_{J-1}^T \end{pmatrix} \right) \\
&= \begin{pmatrix} 1 & \frac{n_1}{n} \\ \frac{n_1}{n} & \frac{n_1}{n} \end{pmatrix} \otimes (\mathbf{D}_{\boldsymbol{\pi}^*(\boldsymbol{\theta}_0)} - \boldsymbol{\pi}^*(\boldsymbol{\theta}_0)\boldsymbol{\pi}^{*T}(\boldsymbol{\theta}_0)),
\end{aligned}$$

and then

$$\mathcal{I}_F(\boldsymbol{\theta}_0) = \begin{pmatrix} 1 & \nu_1 \\ \nu_1 & \nu_1 \end{pmatrix} \otimes (\mathbf{D}_{\boldsymbol{\pi}^*(\boldsymbol{\theta}_0)} - \boldsymbol{\pi}^*(\boldsymbol{\theta}_0)\boldsymbol{\pi}^{*T}(\boldsymbol{\theta}_0)). \quad (28)$$

■

The following theorem establishes that the asymptotic distribution of the families of test statistics (24) and (25) corresponds to a J -dimensional chi-bar squared random variable, a mixture of J chi-squared distributions. Let $E = \{1, \dots, J-1\}$ be the whole set of all row-indices of matrix \mathbf{R} , $\mathcal{F}(E)$ the family of all possible subsets of E , and $\mathbf{R}(S)$ is a submatrix of \mathbf{R} with row-indices belonging to $S \in \mathcal{F}(E)$. We must not forget that $\mathbf{R} = (\mathbf{0}_{(J-1) \times (J-1)}, \mathbf{G}_{J-1})$ and therefore $\mathbf{R}(S) = (\mathbf{0}_{\text{card}(S) \times (J-1)}, \mathbf{G}_{J-1}(S))$.

We denote by $\mathbf{H}(\boldsymbol{\theta})$ the following $(J-1) \times (J-1)$ tridiagonal matrix

$$\mathbf{H}(\boldsymbol{\theta}) = \frac{1}{\nu_1 \nu_2} \begin{pmatrix} \frac{\pi_1(\boldsymbol{\theta}) + \pi_2(\boldsymbol{\theta})}{\pi_1(\boldsymbol{\theta})\pi_2(\boldsymbol{\theta})} & -\frac{1}{\pi_2(\boldsymbol{\theta})} & & & \\ -\frac{1}{\pi_1(\boldsymbol{\theta})} & \frac{\pi_2(\boldsymbol{\theta}) + \pi_3(\boldsymbol{\theta})}{\pi_2(\boldsymbol{\theta})\pi_3(\boldsymbol{\theta})} & -\frac{1}{\pi_3(\boldsymbol{\theta})} & & \\ & -\frac{1}{\pi_3(\boldsymbol{\theta})} & \frac{\pi_3(\boldsymbol{\theta}) + \pi_4(\boldsymbol{\theta})}{\pi_3(\boldsymbol{\theta})\pi_4(\boldsymbol{\theta})} & \ddots & \\ & & \ddots & \ddots & -\frac{1}{\pi_{J-1}(\boldsymbol{\theta})} \\ & & & -\frac{1}{\pi_{J-1}(\boldsymbol{\theta})} & \frac{\pi_{J-1}(\boldsymbol{\theta}) + \pi_J(\boldsymbol{\theta})}{\pi_{J-1}(\boldsymbol{\theta})\pi_J(\boldsymbol{\theta})} \end{pmatrix}, \quad (29)$$

and by $\mathbf{H}(S_1, S_2, \boldsymbol{\theta})$ the submatrix of $\mathbf{H}(\boldsymbol{\theta})$ obtained by deleting from it the row-indices contained in the set S_1 and column-indices contained in the set S_2 .

Theorem 2 Under H_0 , the asymptotic distribution of $S_\phi(\mathbf{p}(\tilde{\boldsymbol{\theta}}), \mathbf{p}(\hat{\boldsymbol{\theta}}))$ and $T_\phi(\bar{\mathbf{p}}, \mathbf{p}(\tilde{\boldsymbol{\theta}}), \mathbf{p}(\hat{\boldsymbol{\theta}}))$ is

$$\lim_{n_1, n_2 \rightarrow \infty} \Pr \left(S_\phi(\mathbf{p}(\tilde{\boldsymbol{\theta}}), \mathbf{p}(\hat{\boldsymbol{\theta}})) \leq x \right) = \lim_{n_1, n_2 \rightarrow \infty} \Pr \left(T_\phi(\bar{\mathbf{p}}, \mathbf{p}(\tilde{\boldsymbol{\theta}}), \mathbf{p}(\hat{\boldsymbol{\theta}})) \leq x \right) = \sum_{j=0}^{J-1} w_j(\boldsymbol{\theta}_0) \Pr \left(\chi_{(J-1)-j}^2 \leq x \right)$$

where $\chi_0^2 = 0$ a.s. and $\{w_j(\boldsymbol{\theta}_0)\}_{j=0}^{J-1}$ is the set of weights such that $\sum_{j=0}^{J-1} w_j(\boldsymbol{\theta}_0) = 1$ and

$$w_j(\boldsymbol{\theta}_0) = \sum_{S \in \mathcal{F}(E), \text{card}(S)=j} \Pr(\mathbf{Z}_1(S) \geq \mathbf{0}_j) \Pr(\mathbf{Z}_2(S) \geq \mathbf{0}_{(J-1)-j}), \quad (30)$$

where

$$\begin{aligned}
\mathbf{Z}_1(S) &\sim \mathcal{N}(\mathbf{0}_{\text{card}(S)}, \mathbf{H}^{-1}(S, S, \boldsymbol{\theta}_0)), \\
\mathbf{Z}_2(S) &\sim \mathcal{N}(\mathbf{0}_{(J-1)-\text{card}(S)}, \mathbf{H}(S^C, S^C, \boldsymbol{\theta}_0) - \mathbf{H}(S^C, S, \boldsymbol{\theta}_0)\mathbf{H}^{-1}(S, S, \boldsymbol{\theta}_0)\mathbf{H}^T(S^C, S, \boldsymbol{\theta}_0)),
\end{aligned}$$

$S^C = E - S$ and $\text{card}(S)$ denotes the cardinal of the set S .

Proof. By following similar arguments of Martín and Balakrishnan we obtain $\mathbf{H}(S, S, \boldsymbol{\theta}_0) = \mathbf{R}(S)\mathcal{I}_F^{-1}(\boldsymbol{\theta}_0)\mathbf{R}^T(S)$ (see Appendix A.3, for the details). In particular, $\mathbf{H}(\boldsymbol{\theta}_0) = \mathbf{H}(S, S, \boldsymbol{\theta}_0)$ with $S = E$, i.e.

$$\begin{aligned}\mathbf{H}(\boldsymbol{\theta}_0) &= \mathbf{R}(E)\mathcal{I}_F^{-1}(\boldsymbol{\theta}_0)\mathbf{R}^T(E) \\ &= (\mathbf{0}_{(J-1) \times (J-1)}, \mathbf{G}_{J-1})\mathcal{I}_F^{-1}(\boldsymbol{\theta}_0)(\mathbf{0}_{(J-1) \times (J-1)}, \mathbf{G}_{J-1})^T,\end{aligned}$$

where $\mathcal{I}_F(\boldsymbol{\theta}_0)$ is (28). By following the properties of the inverse of the Kronecker product for calculating the inverse of (28),

$$\begin{aligned}\mathcal{I}_F^{-1}(\boldsymbol{\theta}_0) &= \begin{pmatrix} 1 & \nu_1 \\ \nu_1 & \nu_1 \end{pmatrix}^{-1} \otimes (\mathbf{D}_{\boldsymbol{\pi}^*(\boldsymbol{\theta}_0)} - \boldsymbol{\pi}^*(\boldsymbol{\theta}_0)\boldsymbol{\pi}^{*T}(\boldsymbol{\theta}_0))^{-1} \\ &= \begin{pmatrix} \frac{1}{\nu_2} & -\frac{1}{\nu_2} \\ -\frac{1}{\nu_2} & \frac{1}{\nu_2} \end{pmatrix} \otimes \left(\mathbf{D}_{\boldsymbol{\pi}^*(\boldsymbol{\theta}_0)}^{-1} + \frac{1}{\pi_J(\boldsymbol{\theta}_0)} \mathbf{1}_{J-1} \mathbf{1}_{J-1}^T \right),\end{aligned}$$

and replacing it in the previous expression of $\mathbf{H}(\boldsymbol{\theta}_0)$,

$$\begin{aligned}\mathbf{H}(\boldsymbol{\theta}_0) &= \frac{1}{\nu_1 \nu_2} \mathbf{G}_{J-1} \left(\mathbf{D}_{\boldsymbol{\pi}^*(\boldsymbol{\theta}_0)}^{-1} + \frac{1}{\pi_J(\boldsymbol{\theta}_0)} \mathbf{1}_{J-1} \mathbf{1}_{J-1}^T \right) \mathbf{G}_{J-1}^T \\ &= \frac{1}{\nu_1 \nu_2} \left(\mathbf{G}_{J-1} \mathbf{D}_{\boldsymbol{\pi}^*(\boldsymbol{\theta}_0)}^{-1} \mathbf{G}_{J-1}^T + \frac{1}{\pi_J(\boldsymbol{\theta}_0)} \mathbf{e}_{J-1} \mathbf{e}_{J-1}^T \right),\end{aligned}$$

which is equal to (29). ■

Even though there is an equality in (18), $\boldsymbol{\theta}$ is not a fixed vector under the null hypothesis since such an equality is effective only for $\boldsymbol{\theta}_{12}$, and thus $\boldsymbol{\theta}_2$ is a vector of nuisance parameters. This means that we have a composite null hypothesis which requires estimation of $\boldsymbol{\theta} \in \Theta_0$, through $\hat{\boldsymbol{\theta}}$ and we cannot use directly the results based on Theorem 2. The tests performed replacing the parameter $\boldsymbol{\theta}_0$ of the asymptotic distribution by $\hat{\boldsymbol{\theta}}$ are called “local tests” (see Dardanoni and Forcina (1998)) and they are usually considered to be good approximations of the theoretical tests.

In relation to the weights, $\{w_j(\boldsymbol{\theta}_0)\}_{j=1,\dots,J}$, there are explicit expressions when $J \in \{2, 3, 4\}$ based on the matrix given in (29) and formulas (3.24), (3.25) and (3.26) in Silvapulle and Sen (2005, page 80). When $J = 2$, $w_0(\boldsymbol{\theta}_0) = w_1(\boldsymbol{\theta}_0) = \frac{1}{2}$. When $J = 3$, the estimators of the weights are

$$\begin{cases} w_0(\hat{\boldsymbol{\theta}}) = \frac{1}{2} - w_2(\hat{\boldsymbol{\theta}}), \\ w_1(\hat{\boldsymbol{\theta}}) = \frac{1}{2}, \\ w_2(\hat{\boldsymbol{\theta}}) = \frac{1}{2\pi} \arccos \hat{\rho}_{12}, \end{cases} \quad (31)$$

where

$$\hat{\rho}_{ij} = \frac{\hat{\sigma}_{ij}}{\sqrt{\hat{\sigma}_{ii}\hat{\sigma}_{jj}}} = -\sqrt{\frac{(N_{1i} + N_{2i})(N_{1,j+1} + N_{2,j+1})}{(N_{1i} + N_{2i} + N_{1j} + N_{2j})(N_{1j} + N_{2j} + N_{1,j+1} + N_{2,j+1})}}, \quad (32)$$

is the correlation associated with the i -th and j -th variable of a central random variable with variance-covariance matrix

$$\mathbf{H}(\hat{\boldsymbol{\theta}}) = \frac{1}{\hat{\nu}_1 \hat{\nu}_2} \begin{pmatrix} \frac{\pi_1(\hat{\boldsymbol{\theta}}) + \pi_2(\hat{\boldsymbol{\theta}})}{\pi_1(\hat{\boldsymbol{\theta}})\pi_2(\hat{\boldsymbol{\theta}})} & -\frac{1}{\pi_2(\hat{\boldsymbol{\theta}})} \\ -\frac{1}{\pi_2(\hat{\boldsymbol{\theta}})} & \frac{\pi_2(\hat{\boldsymbol{\theta}}) + \pi_3(\hat{\boldsymbol{\theta}})}{\pi_2(\hat{\boldsymbol{\theta}})\pi_3(\hat{\boldsymbol{\theta}})} \end{pmatrix},$$

where $\pi_j(\hat{\boldsymbol{\theta}}) = \frac{N_{1j} + N_{2j}}{n}$. When $J = 4$,

$$\begin{cases} w_0(\hat{\boldsymbol{\theta}}) = \frac{1}{4\pi} (2\pi - \arccos \hat{\rho}_{12} - \arccos \hat{\rho}_{13} - \arccos \hat{\rho}_{23}), \\ w_1(\hat{\boldsymbol{\theta}}) = \frac{1}{4\pi} (3\pi - \arccos \hat{\rho}_{12,3} - \arccos \hat{\rho}_{13,2} - \arccos \hat{\rho}_{23,1}), \\ w_2(\hat{\boldsymbol{\theta}}) = \frac{1}{2} - w_0(\hat{\boldsymbol{\theta}}), \\ w_3(\hat{\boldsymbol{\theta}}) = \frac{1}{2} - w_1(\hat{\boldsymbol{\theta}}), \end{cases} \quad (33)$$

which depend on the estimation of the marginal (32) and conditional correlations

$$\hat{\rho}_{ij \cdot k} = \frac{\hat{\rho}_{ij} - \hat{\rho}_{ik} \hat{\rho}_{kj}}{\sqrt{(1 - \hat{\rho}_{ik}^2)(1 - \hat{\rho}_{kj}^2)}},$$

associated with the i -th and j -th variable, given a value of the k -th variable, of a central random variable with variance-covariance matrix

$$\mathbf{H}(\hat{\boldsymbol{\theta}}) = \frac{1}{\hat{\nu}_1 \hat{\nu}_2} \begin{pmatrix} \frac{\pi_1(\hat{\boldsymbol{\theta}}) + \pi_2(\hat{\boldsymbol{\theta}})}{\pi_1(\hat{\boldsymbol{\theta}}) \pi_2(\hat{\boldsymbol{\theta}})} & -\frac{1}{\pi_2(\hat{\boldsymbol{\theta}})} & 0 \\ -\frac{1}{\pi_2(\hat{\boldsymbol{\theta}})} & \frac{\pi_2(\hat{\boldsymbol{\theta}}) + \pi_3(\hat{\boldsymbol{\theta}})}{\pi_2(\hat{\boldsymbol{\theta}}) \pi_3(\hat{\boldsymbol{\theta}})} & -\frac{1}{\pi_3(\hat{\boldsymbol{\theta}})} \\ 0 & -\frac{1}{\pi_3(\hat{\boldsymbol{\theta}})} & \frac{\pi_3(\hat{\boldsymbol{\theta}}) + \pi_4(\hat{\boldsymbol{\theta}})}{\pi_3(\hat{\boldsymbol{\theta}}) \pi_4(\hat{\boldsymbol{\theta}})} \end{pmatrix}.$$

It is interesting to point out that the factor related to the sample size in each multinomial sample, $\frac{1}{\hat{\nu}_1 \hat{\nu}_2}$, have no effect in the expression of estimator for the weights of the chi-bar squared distribution. These formulas will be considered in the forthcoming sections. It is worthwhile to mention that the normal orthant probabilities for the weights given in (30), can also be computed for any value of J using the `mvtnorm` R package (see <http://CRAN.R-project.org/package=mvtnorm>, for details).

5 Numerical example

In this section the data set of the introduction (Table 1), where $J = 4$, is analyzed. The sample, a realization of \mathbf{N} , is summarized in the following vector

$$\mathbf{n} = (n_{11}, n_{12}, n_{13}, n_{14}, n_{21}, n_{22}, n_{23}, n_{24})^T = (12, 10, 4, 6, 5, 8, 8, 11)^T.$$

The order restricted MLE under likelihood ratio order, obtained through the E04UCF subroutine of NAG Fortran library (<http://www.nag.co.uk/numeric/fl/FLdescription.asp>), is

$$\tilde{\boldsymbol{\theta}} = (1.5173, 0.8650, 0, -0.8009, -0.3309, -0.3483, -0.6419)^T.$$

The estimation of the probability vectors of interest is

$$\begin{aligned} \bar{\mathbf{p}} &= (0.1875, 0.1563, 0.0625, 0.0938, 0.0781, 0.1250, 0.1250, 0.1719)^T, \\ \mathbf{p}(\tilde{\boldsymbol{\theta}}) &= (0.1875, 0.1562, 0.0647, 0.0916, 0.0781, 0.1250, 0.1228, 0.1740)^T, \\ \mathbf{p}(\hat{\boldsymbol{\theta}}) &= (0.1328, 0.1406, 0.0938, 0.1328, 0.1328, 0.1406, 0.0938, 0.1328)^T, \end{aligned}$$

and the estimation of the weights, based on (33), are

$$w_0(\hat{\boldsymbol{\theta}}) = 0.038155, \quad w_1(\hat{\boldsymbol{\theta}}) = 0.242013, \quad w_2(\hat{\boldsymbol{\theta}}) = 0.461845, \quad w_3(\hat{\boldsymbol{\theta}}) = 0.257987.$$

In order to solve analytically the example we shall consider a particular function ϕ in (24) and (25). Taking

$$\phi_\lambda(x) = \frac{x^\lambda - x - \lambda(x - 1)}{\lambda(\lambda + 1)},$$

we get the “the power divergence family”

$$d_{\phi_\lambda}(\mathbf{p}, \mathbf{q}) = \frac{1}{\lambda(\lambda + 1)} \left(\sum_{i=1}^2 \sum_{j=1}^J \frac{p_{ij}^{\lambda+1}}{q_{ij}^\lambda(\hat{\boldsymbol{\theta}})} - 1 \right)$$

in such a way that for each $\lambda \in \mathbb{R} - \{-1, 0\}$ a different divergence measure is obtained, and thus

$$T_\lambda = T_{\phi_\lambda}(\bar{\mathbf{p}}, \mathbf{p}(\tilde{\boldsymbol{\theta}}), \mathbf{p}(\hat{\boldsymbol{\theta}})) = \frac{2n}{\lambda(\lambda+1)} \left(\sum_{i=1}^2 \sum_{j=1}^J \frac{\bar{p}_{ij}^{\lambda+1}}{p_{ij}^\lambda(\hat{\boldsymbol{\theta}})} - \sum_{i=1}^2 \sum_{j=1}^J \frac{\bar{p}_{ij}^{\lambda+1}}{p_{ij}^\lambda(\tilde{\boldsymbol{\theta}})} \right), \quad (34)$$

$$S_\lambda = S_{\phi_\lambda}(\mathbf{p}(\tilde{\boldsymbol{\theta}}), \mathbf{p}(\hat{\boldsymbol{\theta}})) = \frac{2n}{\lambda(\lambda+1)} \left(\sum_{i=1}^2 \sum_{j=1}^J \frac{p_{ij}^{\lambda+1}(\tilde{\boldsymbol{\theta}})}{p_{ij}^\lambda(\hat{\boldsymbol{\theta}})} - 1 \right). \quad (35)$$

It is also possible to cover the real line for λ , by defining

$$d_{\phi_\lambda}(\mathbf{p}, \mathbf{q}) = \lim_{\ell \rightarrow \lambda} d_{\phi_\ell}(\mathbf{p}, \mathbf{q}), \quad \lambda \in \{-1, 0\},$$

and by considering $T_\lambda = \lim_{\lambda \rightarrow \ell} T_\ell$, $S_\lambda = \lim_{\lambda \rightarrow \ell} S_\ell$, for $\lambda \in \{0, -1\}$, i.e.

$$T_0 = T_{\phi_0}(\bar{\mathbf{p}}, \mathbf{p}(\tilde{\boldsymbol{\theta}}), \mathbf{p}(\hat{\boldsymbol{\theta}})) = G^2 = 2n \sum_{i=1}^2 \sum_{j=1}^J \bar{p}_{ij} \log \frac{p_{ij}(\tilde{\boldsymbol{\theta}})}{p_{ij}(\hat{\boldsymbol{\theta}})}, \quad (36)$$

$$T_{-1} = T_{\phi_{-1}}(\bar{\mathbf{p}}, \mathbf{p}(\tilde{\boldsymbol{\theta}}), \mathbf{p}(\hat{\boldsymbol{\theta}})) = 2n \left(\sum_{i=1}^2 \sum_{j=1}^J p_{ij}(\hat{\boldsymbol{\theta}}) \log \frac{p_{ij}(\hat{\boldsymbol{\theta}})}{\bar{p}_{ij}} - \sum_{i=1}^2 \sum_{j=1}^J p_{ij}(\tilde{\boldsymbol{\theta}}) \log \frac{p_{ij}(\tilde{\boldsymbol{\theta}})}{\bar{p}_{ij}} \right) \quad (37)$$

and

$$S_0 = S_{\phi_0}(\mathbf{p}(\tilde{\boldsymbol{\theta}}), \mathbf{p}(\hat{\boldsymbol{\theta}})) = 2nd_{Kull}(\mathbf{p}(\tilde{\boldsymbol{\theta}}), \mathbf{p}(\hat{\boldsymbol{\theta}})) = 2n \sum_{i=1}^2 \sum_{j=1}^J p_{ij}(\tilde{\boldsymbol{\theta}}) \log \frac{p_{ij}(\tilde{\boldsymbol{\theta}})}{p_{ij}(\hat{\boldsymbol{\theta}})}, \quad (38)$$

$$S_{-1} = S_{\phi_{-1}}(\mathbf{p}(\tilde{\boldsymbol{\theta}}), \mathbf{p}(\hat{\boldsymbol{\theta}})) = 2nd_{Kull}(\mathbf{p}(\hat{\boldsymbol{\theta}}), \mathbf{p}(\tilde{\boldsymbol{\theta}})) = 2n \sum_{j=1}^J p_{ij}(\hat{\boldsymbol{\theta}}) \log \frac{p_{ij}(\hat{\boldsymbol{\theta}})}{p_{ij}(\tilde{\boldsymbol{\theta}})}. \quad (39)$$

It is well known that $d_{\phi_0}(\mathbf{p}, \mathbf{q}) = d_{Kull}(\mathbf{p}, \mathbf{q})$ and $d_{\phi_1}(\mathbf{p}, \mathbf{q}) = d_{Pearson}(\mathbf{p}, \mathbf{q})$, which is very interesting since G^2 and X^2 are members of the power divergence based test-statistics. It is also worthwhile to mention that $d_{\phi_{-1}}(\mathbf{p}, \mathbf{q}) = d_{Kull}(\mathbf{q}, \mathbf{p})$.

In Table 3, the power divergence based test-statistics for some values of λ in $\Lambda = \{-1.5, -1, -\frac{1}{2}, 0, \frac{2}{3}, 1, 1.5, 2\}$, and their corresponding asymptotic p -values are shown. In all of them it is concluded, with a significance level equal to 0.05, that an equal effect of both treatments is rejected and hence the treatment is more effective than the control to heal the ulcer.

test-statistic	$\lambda = -1.5$	$\lambda = -1$	$\lambda = -\frac{1}{2}$	$\lambda = 0$	$\lambda = \frac{2}{3}$	$\lambda = 1$	$\lambda = 1.5$	$\lambda = 2$	$\lambda = 3$
T_λ	6.5323	6.3215	6.1562	6.0323	5.9261	5.8965	5.8803	5.8965	6.0244
$p\text{-value}(T_\lambda)$	0.0175	0.0194	0.0211	0.0225	0.0238	0.0241	0.0243	0.0241	0.0226
S_λ	6.5277	6.3189	6.1551	6.0323	5.9270	5.8977	5.8815	5.8977	6.0244
$p\text{-value}(S_\lambda)$	0.0175	0.0195	0.0212	0.0225	0.0238	0.0241	0.0243	0.0241	0.0226

Table 3: Power divergence based test-statistics and asymptotic p -values for the data given Table 1.

The p -values given in Table 3 were obtained by the following algorithm:
Let $T \in \{T_\lambda, S_\lambda\}_{\lambda \in \Lambda}$ be the test-statistic associated with (10). In the following steps the corresponding asymptotic p -value, based on the asymptotic distribution of Theorem 2, is calculated once it is suppose we have $\{w_j(\hat{\boldsymbol{\theta}})\}_{j=0}^{J-1}$:

STEP 1: Using n calculate $p(\hat{\theta})$ taking into account (19).
STEP 2: Using $p(\hat{\theta})$ calculate value t of test-statistic T using the corresponding expression in (34)–(39).
STEP 3: If $T \leq 0$ then compute $p\text{-value}(T) := 1$ and STOP, otherwise compute $p\text{-value}(T) := 0$.
STEP 4: For $j = 0, \dots, J - 2$, do $p\text{-value}(T) := p\text{-value}(T) + w_j(\hat{\theta}) \Pr(\chi_{(J-1)-j}^2 > t)$.
E.g., the NAG Fortran library subroutine G01ECF can be useful.

6 Simulation study

In this Section the performance of the power divergence test statistics (34)–(39) is studied in terms of the simulated exact size and simulated power of the test, based on small and moderate sample sizes. A simulation experiment with four scenarios is designed in Table 4, taking into account the sample sizes of the two independent samples. With respect to the choice of λ , the parameters for the power divergence test statistics, the interest is focused on the interval $[-1.5, 3]$. Note that the test-statistics applied in the numerical example are covered as particular cases.

scenarios	sc. A	sc. B	sc. C	sc. D
n_1	10	12	18	24
n_2	13	16	24	32

Table 4: Scenarios of the simulation study based on sample sizes.

The algorithm described in Section 5 is taken into account to calculate the p -value of each test-statistic $T \in \{T_\lambda, S_\lambda\}_{\lambda \in [-1.5, 3]}$, with a sample \mathbf{N} , and this is repeated independently $R = 25\,000$ times. The simulated exact power was computed as

$$\hat{\beta}_T = \hat{\beta}_T(\delta) = \frac{\text{number of replications of } T \text{ for which the } p\text{-value is less than } \alpha}{R},$$

for the probability vectors

$$\begin{aligned} \boldsymbol{\pi}_i(\boldsymbol{\theta}(\delta)) &= (\pi_{i1}(\boldsymbol{\theta}(\delta)), \pi_{i2}(\boldsymbol{\theta}(\delta)), \pi_{i3}(\boldsymbol{\theta}(\delta)))^T \\ \pi_{ij}(\boldsymbol{\theta}(\delta)) &= \frac{1}{3} \frac{1 + i(j-1)\delta}{1 + i\delta}, \quad i = 1, 2, \quad j = 1, 2, 3, \end{aligned}$$

for $\delta \in \Xi = \{0.1, 0.5, 1.0, 1.5\}$. The simulated exact size was computed as

$$\hat{\alpha}_T = \frac{\text{number of replications of } T \text{ for which the } p\text{-value is less than } \alpha}{R},$$

for the probability vectors

$$\begin{aligned} \boldsymbol{\pi}_i(\boldsymbol{\theta}_0) &= (\pi_{i1}(\boldsymbol{\theta}_0), \pi_{i2}(\boldsymbol{\theta}_0), \pi_{i3}(\boldsymbol{\theta}_0))^T \\ \pi_{ij}(\boldsymbol{\theta}_0) &= \frac{1}{3}, \quad i = 1, 2, \quad j = 1, 2, 3, \end{aligned}$$

which corresponds to the case of $\delta = 0$ for $\boldsymbol{\pi}_i(\boldsymbol{\theta}(\delta))$.

In Table 5 the local odds ratios,

$$\vartheta_j = \vartheta_j(\delta) = \frac{1 + (j-1)\delta}{1 + 2(j-1)\delta} \frac{1 + 2j\delta}{1 + j\delta},$$

$j = 1, 2$, are shown for $\delta \in \{0\} \cup \Xi$. Notice that in $\boldsymbol{\vartheta} = \boldsymbol{\vartheta}(\delta) = (\vartheta_1(\delta), \vartheta_2(\delta))^T$ some of the components are further from $\boldsymbol{\vartheta}(0) = \mathbf{1}_2$ (null hypothesis), as the value of $\delta > 0$ is further from 0. This means that a greater value of the estimation of the power function might be obtained, as $\delta > 0$ is greater. This claim is supported by the fact that some values of the components of $\boldsymbol{\vartheta} = \boldsymbol{\vartheta}(\delta)$ decrease as $\delta > 0$ increases but more slowly than the others increase. In addition, for a fixed value of $\delta > 0$, it is expected a greater value of $\hat{\beta}_T(\delta)$, as n is greater (the worst powers in Scenario A and the best powers in Scenario D). We have also added in Table 5 the last three rows for two reasons, first, to show that for any fixed value of δ , $\pi_{2j}(\boldsymbol{\theta}(\delta))/\pi_{1j}(\boldsymbol{\theta}(\delta))$ is non-decreasing as j , the ordinal category, increases and second, to clarify the meaning of the two asterisks contained in the table. It is clear that for a big value of δ , $\pi_{i1}(\boldsymbol{\theta}(\delta)) > 0$ goes to zero on the right for $i = 1, 2$, but in the practice, due to the empty cells in the contingency table, the estimator of the ratio $\pi_{21}(\boldsymbol{\theta}(\delta))/\pi_{11}(\boldsymbol{\theta}(\delta))$ becomes 1 rather than $\frac{1}{2}$ (and $\vartheta_1(\delta)$ becomes 1). This was our experience when we used values of δ bigger than 1.5, i.e. the power becomes quite little in the practice.

	$\delta = 0$	$\delta = 0.1$	$\delta = 0.5$	$\delta = 1$	$\delta = 1.5$	$\delta = \infty$
$\vartheta_1 = \vartheta_1(\delta)$	1.000	1.091	1.333	1.500	1.600	2.00*
$\vartheta_2 = \vartheta_2(\delta)$	1.000	1.069	1.125	1.111	1.094	1.00
$\pi_{21}(\boldsymbol{\theta}(\delta))/\pi_{11}(\boldsymbol{\theta}(\delta))$	0.33/0.33	0.28/0.30	0.17/0.22	0.11/0.17	0.08/0.13	0.50*
$\pi_{22}(\boldsymbol{\theta}(\delta))/\pi_{12}(\boldsymbol{\theta}(\delta))$	0.33/0.33	0.33/0.33	0.33/0.33	0.33/0.33	0.33/0.33	1.00

Table 5: Theoretical local odd ratios for the Monte Carlo study.

Once a nominal size $\alpha = 0.05$ is established, Table 6 summarizes the simulated exact sizes in all the scenarios for the test-statistic $T \in \{T_\lambda, S_\lambda\}_{\lambda \in \Lambda}$, with $\Lambda = \{-1.5, -1, -\frac{1}{2}, 0, \frac{2}{3}, 1, 1.5, 2, 3\}$. We have plotted 3×2 graphs in Figures 1-4 and we refer them as plots in three rows. In the first row of Figures 1-4 we can see on the left the exact power in all the scenarios for the test-statistic $\{T_\lambda\}_{\lambda \in [-1.5, 3]}$ and on the right for the test-statistic $\{S_\lambda\}_{\lambda \in [-1.5, 3]}$. In order to make a comparison of exact powers, we cannot directly proceed without considering the exact sizes. For this reason we are going to give a procedure based on two steps.

Step 1: We are going to check for all the power divergence based test-statistics the criterion given by Dale (1986), i.e.,

$$|\text{logit}(1 - \hat{\alpha}_T) - \text{logit}(1 - \alpha)| \leq e \quad (40)$$

with $\text{logit}(p) = \log\left(\frac{p}{1-p}\right)$. We only consider the values of λ such that $\hat{\alpha}_T$ verifies (40) with $e = 0.35$, then we shall only consider the test-statistics such that $\hat{\alpha}_T \in [0.0357, 0.0695]$, in all the scenarios. This criterion has been considered for some authors, see for instance Cressie et al. (2003) and Martín and Pardo (2012). The cases satisfying the criterion are marked in bold in Table 6, and comprise those values in the abscissa of the plot between the dashed band (the dashed line in the middle represents the nominal size), and we can conclude that we must not consider in our study $T \in \{T_\lambda, S_\lambda\}_{\lambda \in [-1.5, -0.4]}$.

Step 2: We compare all the test statistics obtained in Step 1 with the classical likelihood ratio test ($G^2 = T_0$) as well as the classical Pearson test statistic ($X^2 = S_1$). To do so, we have calculated the relative local efficiencies

$$\hat{\rho}_T = \hat{\rho}_T(\delta) = \frac{(\hat{\beta}_T(\delta) - \hat{\alpha}_T) - (\hat{\beta}_{T_0}(\delta) - \hat{\alpha}_{T_0})}{\hat{\beta}_{T_0}(\delta) - \hat{\alpha}_{T_0}}, \quad \hat{\rho}_T^* = \hat{\rho}_T^*(\delta) = \frac{(\hat{\beta}_T(\delta) - \hat{\alpha}_T) - (\hat{\beta}_{S_1}(\delta) - \hat{\alpha}_{S_1})}{\hat{\beta}_{S_1}(\delta) - \hat{\alpha}_{S_1}}.$$

In Figures 1-4 the powers and the relative local efficiencies are summarized. The second rows of the figures represent $\hat{\rho}_T$, while in the third row is plotted $\hat{\rho}_T^*$, on the left it is considered $T = T_\lambda$ and $T = S_\lambda$ on the right.

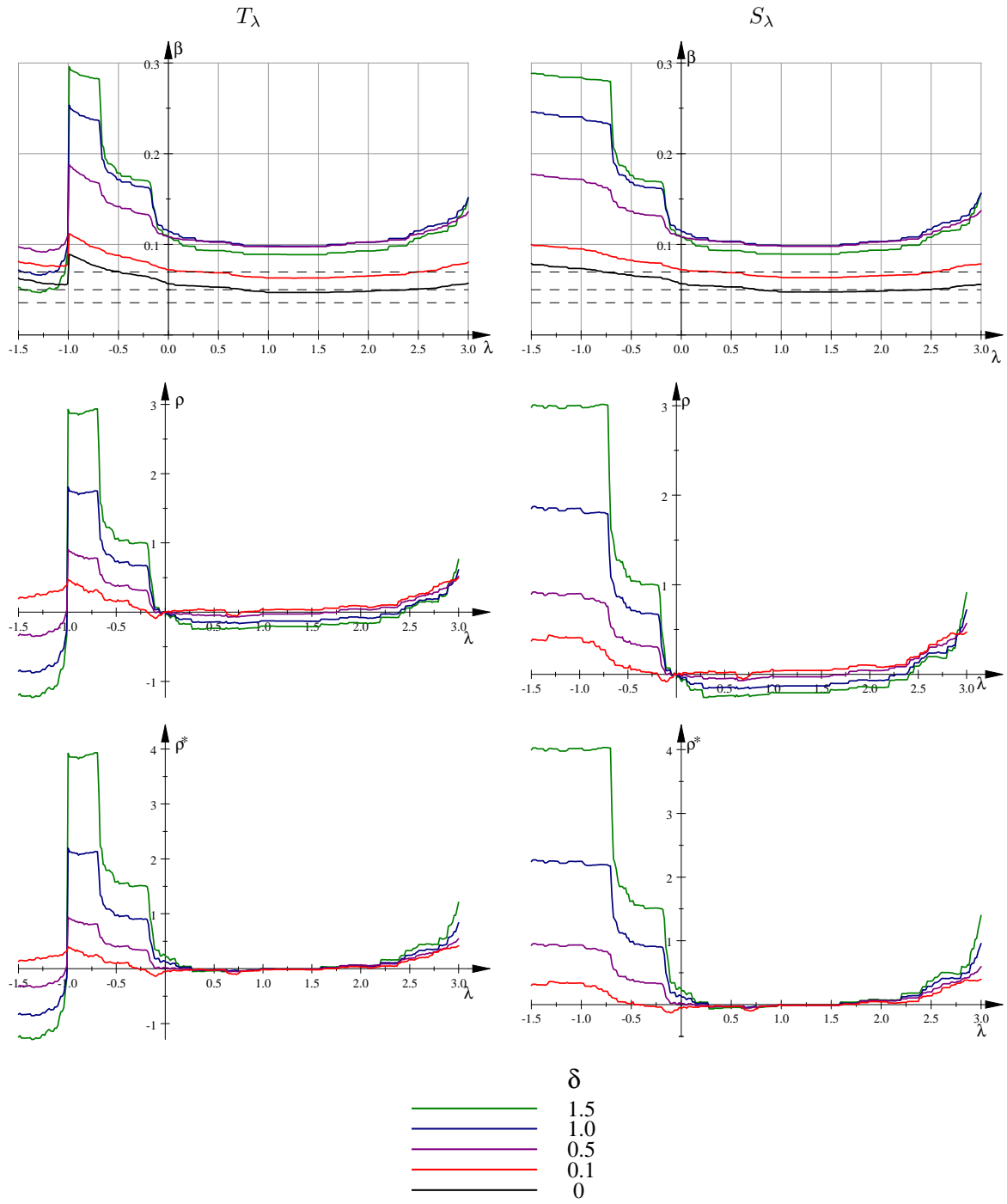


Figure 1: Power and relative local efficiencies for T_λ and S_λ in scenario A.

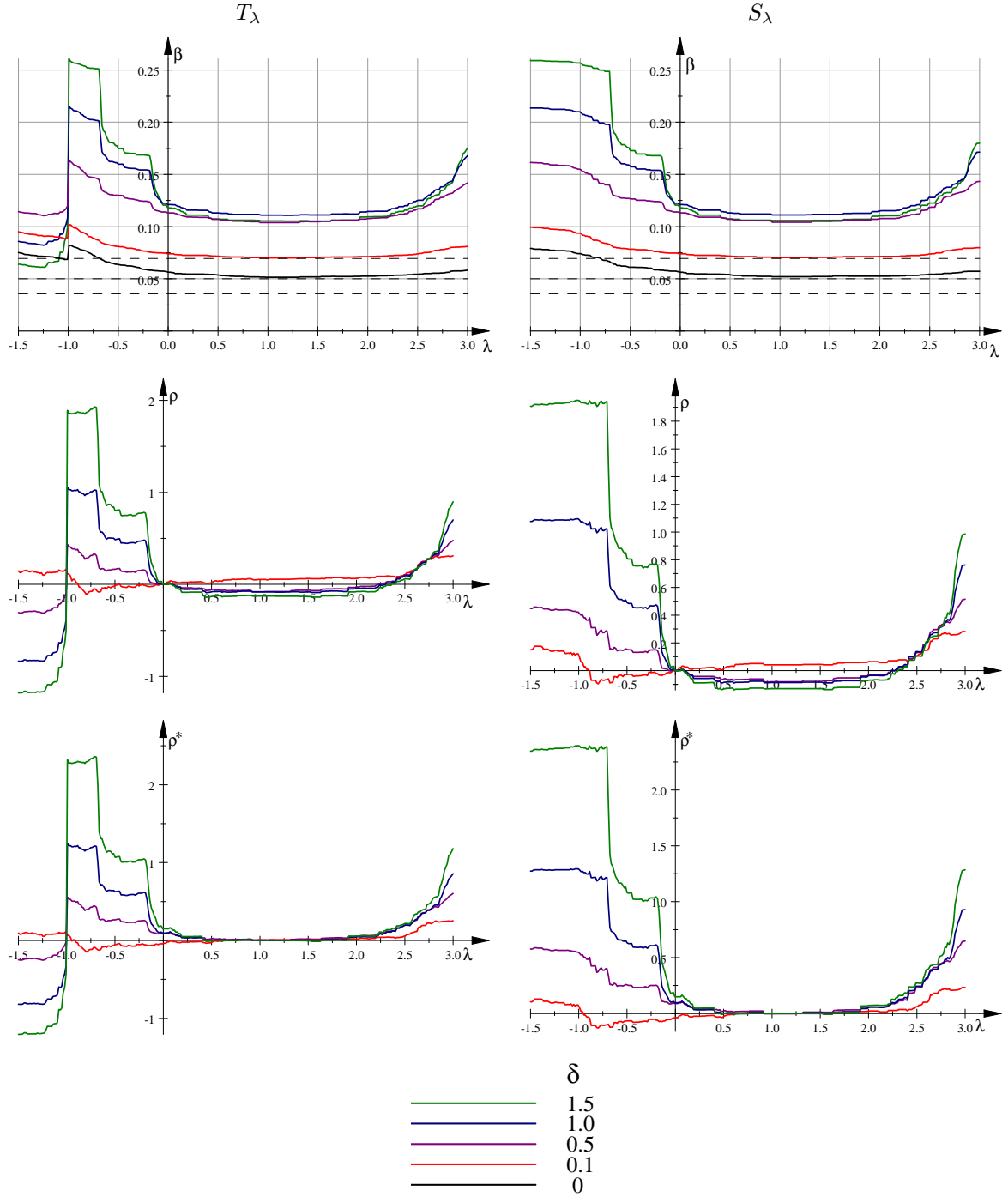


Figure 2: Power and relative local efficiencies for T_λ and S_λ in scenario B.

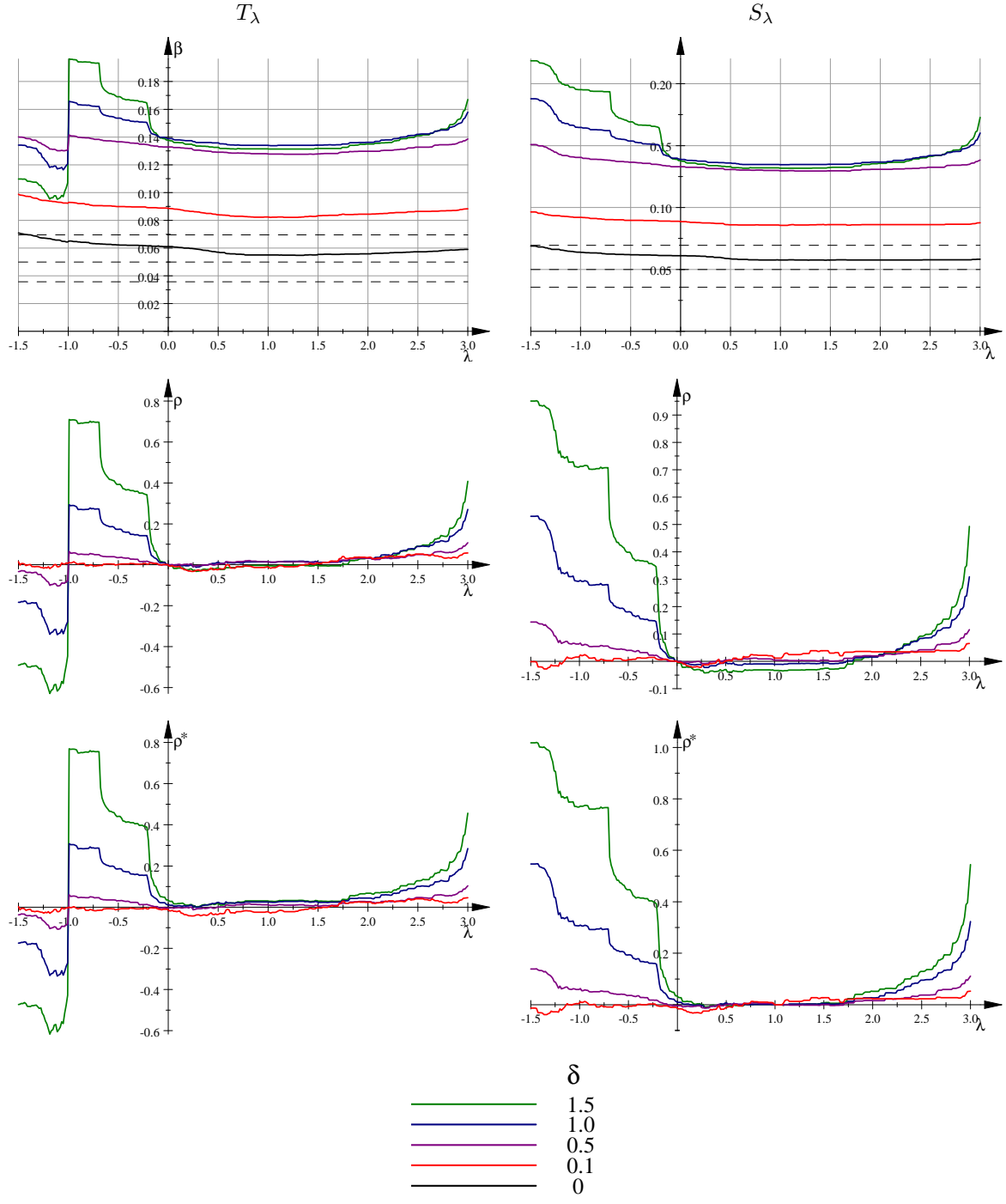


Figure 3: Power and relative local efficiencies for T_λ and S_λ in scenario C.

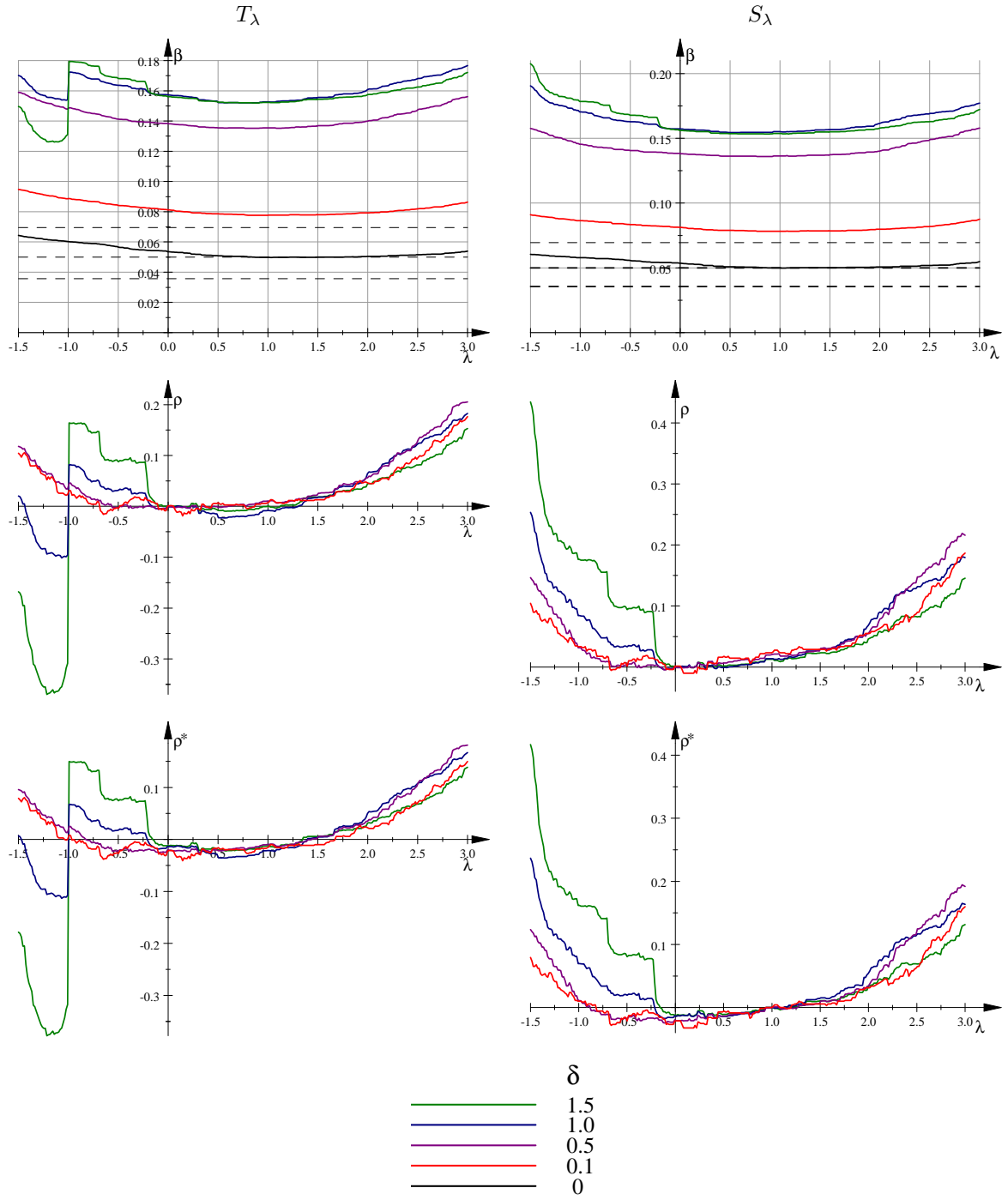


Figure 4: Power and relative local efficiencies for T_λ and S_λ in scenario D.

sc	$\hat{\alpha}_{T_{-1.5}}$	$\hat{\alpha}_{T_{-1}}$	$\hat{\alpha}_{T_{-1/2}}$	$\hat{\alpha}_{T_0}$	$\hat{\alpha}_{T_{2/3}}$	$\hat{\alpha}_{T_1}$	$\hat{\alpha}_{T_{1.5}}$	$\hat{\alpha}_{T_2}$	$\hat{\alpha}_{T_3}$
A	0.0629	0.0540	0.0701	0.0565	0.0512	0.0470	0.0470	0.0484	0.0571
B	0.0752	0.0672	0.0638	0.0563	0.0525	0.0516	0.0518	0.0526	0.0582
C	0.0708	0.0644	0.0623	0.0611	0.0558	0.0551	0.0555	0.0558	0.0591
D	0.0643	0.0602	0.0566	0.0536	0.0505	0.0498	0.0500	0.0504	0.0539
sc	$\hat{\alpha}_{S_{-1.5}}$	$\hat{\alpha}_{S_{-1}}$	$\hat{\alpha}_{S_{-1/2}}$	$\hat{\alpha}_{S_0}$	$\hat{\alpha}_{S_{2/3}}$	$\hat{\alpha}_{S_1}$	$\hat{\alpha}_{S_{1.5}}$	$\hat{\alpha}_{S_2}$	$\hat{\alpha}_{S_3}$
A	0.0789	0.0733	0.0661	0.0565	0.0515	0.0474	0.0474	0.0486	0.0557
B	0.0791	0.0741	0.0612	0.0563	0.0528	0.0520	0.0522	0.0526	0.0572
C	0.0688	0.0640	0.0618	0.0611	0.0580	0.0576	0.0576	0.0576	0.0583
D	0.0605	0.0579	0.0557	0.0536	0.0506	0.0500	0.0501	0.0507	0.0548

Table 6: $\hat{\alpha}_T$, for $T \in \{T_\lambda, S_\lambda\}_{\lambda \in \Lambda}$ in scenarios of Table 4.

The plots are interpreted as follows:

- a) In all the scenarios a similar pattern is observed when plotting the exact power, $\hat{\beta}_T$, for $\lambda \in [-0.6, 3]$ since a U shaped curve is obtained. This means that the exact power is higher in the corners of the interval in comparison with the classical likelihood ratio test ($G^2 = T_0$) as well as the classical Pearson test statistic ($X^2 = S_1$), contained in the middle.
- b) If we pay attention on the local efficiencies with respect to G^2 and X^2 , $\hat{\rho}_T$ and $\hat{\rho}_T^*$, to find positive values of them we need to consider $\lambda \in [-0.6, 0] \cup [2.25, 3]$ and thus it corfims what was said in a). In addition $\hat{\rho}_T^*$ tends to be higher than $\hat{\rho}_T$. This means the test statistics are better with respect to X^2 , in other words, G^2 is better than X^2 . On the other hand, comparing the left hand ($T = T_\lambda$) side of $\hat{\rho}_T$ with the right side ($T = S_\lambda$) and doing the same for $\hat{\rho}_T^*$, a slightly higher values of the local efficiencies of S_λ are seen in comparison with T_λ . For this reason we consider that $\{S_\lambda\}_{\lambda \in [-0.6, 0] \cup [2.25, 3]}$ have a better performance than the classical test-statistics, G^2 and X^2 .
- c) As expected, using asymptotic distribution of test-statistics for comparing the performance for small and moderate sample sizes, the gain in power measured through the local efficiencies is better in scenarios A and B in comparison with scenarios C and D. What is not so common in comparison with usual models of categorical data is to find small size sample sizes with so good performance in exact size as it happens in the current model.

7 Concluding remark

The likelihood ratio ordering is a useful technique for comparing treatments in clinical trials, for this reason it is vitally important to provide test-statistics to improve the classical ones. Having considered an asymptotic distribution for two order restricted treatments, the weights needed to manage the associated asymptotic chi-bar distribution are calculated in a simple way and the useful matrix for that, $\mathbf{H}(\hat{\theta})$, has an easy interpretation in terms of log-linear modeling. The simulation study highlights the good performance of the all the proposed tests in relation to the exact size and the comparison is made in terms of the power. We think that this is a specific characteristic of the likelihood ordering, and this is the reason of having obtained as the best test-statistics a set of values of $\lambda \in [-0.6, 0] \cup [2.25, 3]$ not very common in the literature of phi-divergence test-statistics. As

exception, notice that

$$\begin{aligned}
S_{-1/2} &= S_{d_{\phi_{-1/2}}}(\mathbf{p}(\tilde{\boldsymbol{\theta}}), \mathbf{p}(\hat{\boldsymbol{\theta}})) = 8n \left(1 - \sum_{i=1}^2 \sum_{j=1}^J p_{ij}^{\frac{1}{2}}(\tilde{\boldsymbol{\theta}}) p_{ij}^{\frac{1}{2}}(\hat{\boldsymbol{\theta}}) \right) \\
&= 4n \sum_{i=1}^2 \sum_{j=1}^J \left(p_{ij}^{\frac{1}{2}}(\tilde{\boldsymbol{\theta}}) - p_{ij}^{\frac{1}{2}}(\hat{\boldsymbol{\theta}}) \right)^2 \\
&= 4n \text{Hel}^2(\mathbf{p}(\tilde{\boldsymbol{\theta}}), \mathbf{p}(\hat{\boldsymbol{\theta}})),
\end{aligned}$$

where

$$\text{Hel}(\mathbf{p}(\tilde{\boldsymbol{\theta}}), \mathbf{p}(\hat{\boldsymbol{\theta}})) = \left(\sum_{i=1}^2 \sum_{j=1}^J \left(p_{ij}^{\frac{1}{2}}(\tilde{\boldsymbol{\theta}}) - p_{ij}^{\frac{1}{2}}(\hat{\boldsymbol{\theta}}) \right)^2 \right)^{\frac{1}{2}},$$

is the Hellinger distance between the probability vectors $\mathbf{p}(\tilde{\boldsymbol{\theta}})$ and $\mathbf{p}(\hat{\boldsymbol{\theta}})$. Therefore, one of the test-statistic we are proposing in this paper is a function of the well-known Hellinger distance, which has been used in many different statistical problems.

References

- [1] BARLOW, R. E., BARTHOLOMEW, D. J. AND BRUNK, H.D. (1972). *Statistical inference under order restrictions*. Wiley.
- [2] BAZARAA, M. S., SHERALI, H. D. AND SHETTY, C. M. (2006). *Nonlinear Programming: Theory and Algorithms* (3rd Edition). John Wiley and Sons.
- [3] CHRISTENSEN, R. (1997). *Log-linear models and logistic regression*. Springer.
- [4] CRESSIE, N. AND PARDO, L. (2002). Phi-divergence statistics. *Encyclopedia of Environmetrics* (A. H. Elshaarawi and W. W. Piegorsch, Eds.). Volume 3, 1551-1555, John Wiley and Sons, New York.
- [5] CRESSIE, N. AND PARDO, L. (2003). Minimum phi-divergence estimator and hierarchical testing in log-linear models. *Statistica Sinica*, **10**, 867-884.
- [6] CRESSIE, N., PARDO, L. AND PARDO, M.C. (2003). Size and power considerations for testing loglinear models using ϕ -divergence test statistics. *Statistica Sinica*, **13**, 550-570.
- [7] DALE, J.R. (1986). Asymptotic normality of goodness-of-fit statistics for sparse product multinomials. *Journal of the Royal Statistical Society*, **B**, 48-59.
- [8] DYKSTRA, R. L., KOCBAR, S. AND ROBERTSON, T. (1995). Inference for Likelihood Ratio Ordering in the Two-Sample Problem. *Journal of the American Statistical Association*, **90**, 1034-1040.
- [9] HARVILLE, D. A. (2008). *Matrix algebra from a statistician's perspective*. Springer.
- [10] FERGUSON, T. S. (1996). *A Course in Large Sample Theory*. Chapman & Hall.
- [11] KUDÔ, A. (1963). A multivariate analogue of the one-sided test. *Biometrika*, **50**, 403-418.
- [12] LANG, J. B. (1996). On the Comparison of Multinomial and Poisson Log-Linear Models. *Journal of the Royal Statistical Society Series B*, **58**, 253-266.

- [13] LETIERCE, A., TUBERT-BITTER, P., KRAMAR, A. AND MACCARIO, J. (2003). Two-treatment comparison based on joint toxicity and efficacy ordered alternatives in cancer trials. *Statistics in Medicine*, **22**, 859–868.
- [14] MARTIN, N. AND PARDO, L. (2006). Choosing the best phi-divergence goodness-of-fit statistic in multinomial sampling for loglinear models with linear constraints. *Kybernetika*, **42**, 711–722.
- [15] MARTIN, N. AND PARDO, L. (2008a). New families of estimators and test statistics in log-linear models. *Journal of Multivariate Analysis*, **99**(8), 1590–1609.
- [16] MARTIN, N. AND PARDO, L. (2008b). Phi-divergence estimators for loglinear models with linear constraints and multinomial sampling. *Statistical Papers*, **49**, 15–36.
- [17] MARTIN, N. AND PARDO, L. (2011). Fitting DNA sequences through log-linear modelling with linear constraints. *Statistics: A Journal of Theoretical and Applied Statistics*, **45**, 605–621.
- [18] MARTIN, N. AND PARDO, L. (2012). Poisson-loglinear modeling with linear constraints on the expected cell frequencies. *Sankhya B*, **74**(2), 238–267.
- [19] PARDO, L. (2006). *Statistical Inference Based on Divergence Measures*. Statistics: series of Textbooks and Monographs. Chapman & Hall / CRC.
- [20] SEN, P. K., SINGER, J. M. AND PEDROSO DE LIMA, A. C. (2010). *From Finite Sample to Asymptotic Methods in Statistics*. Cambridge University Press.
- [21] SHAPIRO, A. (1985). Asymptotic Distribution of Test Statistics in the Analysis of Moment Structures Under Inequality Constraints. *Biometrika*, **72**, 133–144.
- [22] SHAPIRO, A. (1988). Toward a Unified Theory of Inequality Constrained Testing in Multivariate Analysis. *International Statistical Review*, **56**, 49–62.
- [23] SILVAPULLE, M. J. AND SEN, P. K. (2005). *Constrained statistical inference. Inequality, order, and shape restrictions*. Wiley Series in Probability and Statistics. Wiley-Interscience (John Wiley & Sons).
- [24] Zografos, K., Ferentinos, K. and Papaioannou, T. (1990). ϕ -divergence statistics: Sampling properties and multinomial goodness of fit and divergence tests. *Communications in Statistics-Theory and Methods*, **19**, 1785–1802.

A Appendix

Suppose we are interested in testing $H_0: \mathbf{R}_{12}\boldsymbol{\theta}_{12} = \mathbf{0}_{J-1}$ vs $H_1: \mathbf{R}_{12}(S)\boldsymbol{\theta}_{12} = \mathbf{0}_{\text{card}(S)}$ and $\mathbf{R}_{12}\boldsymbol{\theta}_{12} \neq \mathbf{0}_{J-1}$. With the complete notation, our interest is,

$$H_0: \mathbf{R}\boldsymbol{\theta} = \mathbf{0}_{J-1} \quad \text{vs} \quad H_1: \mathbf{R}(S)\boldsymbol{\theta} = \mathbf{0}_{\text{card}(S)} \quad \text{and} \quad \mathbf{R}\boldsymbol{\theta} \neq \mathbf{0}_{J-1}. \quad (41)$$

Under H_0 , the parameter space is $\Theta_0 = \{\boldsymbol{\theta} \in \mathbb{R}^{J+1} : \mathbf{R}\boldsymbol{\theta} = \mathbf{0}_{J-1}\}$ and the MLE of $\boldsymbol{\theta}$ in Θ_0 is given by $\hat{\boldsymbol{\theta}} = \arg \max_{\boldsymbol{\theta} \in \Theta_0} \ell(\mathbf{N}; \boldsymbol{\theta})$. Under the alternative hypothesis the parameter space is $\Theta(S) - \Theta_0$, where $\Theta(S) = \{\boldsymbol{\theta} \in \mathbb{R}^{J+1} : \mathbf{R}(S)\boldsymbol{\theta} = \mathbf{0}_{J-1}\}$, that is, under both hypotheses, H_0 and H_1 , the parameter space is $\Theta(S) = \{\boldsymbol{\theta} \in \mathbb{R}^{J+1} : \mathbf{R}(S)\boldsymbol{\theta} = \mathbf{0}_{J-1}\}$ and the MLE of $\boldsymbol{\theta}$ in $\Theta(S)$ is $\hat{\boldsymbol{\theta}}(S) = \arg \max_{\boldsymbol{\theta} \in \Theta(S)} \ell(\mathbf{N}; \boldsymbol{\theta})$. By following the same idea we used for building test-statistics (24)-(25) we shall consider two family of test-statistics based on ϕ -divergence measures,

$$T_\phi(\bar{\mathbf{p}}, \mathbf{p}(\hat{\boldsymbol{\theta}}(S)), \mathbf{p}(\hat{\boldsymbol{\theta}})) = 2n(d_\phi(\bar{\mathbf{p}}, \mathbf{p}(\hat{\boldsymbol{\theta}})) - d_\phi(\bar{\mathbf{p}}, \mathbf{p}(\hat{\boldsymbol{\theta}}(S)))) \quad (42)$$

and

$$S_\phi(\mathbf{p}(\hat{\boldsymbol{\theta}}(S)), \mathbf{p}(\hat{\boldsymbol{\theta}})) = 2nd_\phi(\mathbf{p}(\hat{\boldsymbol{\theta}}(S)), \mathbf{p}(\hat{\boldsymbol{\theta}})). \quad (43)$$

A.1 Proposition

Under H_0 ,

$$S_\phi(\mathbf{p}(\hat{\boldsymbol{\theta}}(S)), \mathbf{p}(\hat{\boldsymbol{\theta}})) = T_\phi(\bar{\mathbf{p}}, \mathbf{p}(\hat{\boldsymbol{\theta}}(S)), \mathbf{p}(\hat{\boldsymbol{\theta}})) + o_p(1), \quad (44)$$

the asymptotic distribution of (42) and (43) is χ_{df}^2 with $df = J - 1 - \text{card}(S)$.

Proof. The second order Taylor expansion of function $d_\phi(\boldsymbol{\theta}) = d_\phi(\mathbf{p}(\boldsymbol{\theta}), \mathbf{p}(\hat{\boldsymbol{\theta}}))$ about $\hat{\boldsymbol{\theta}}$ is

$$d_\phi(\boldsymbol{\theta}) = d_\phi(\hat{\boldsymbol{\theta}}) + (\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})^T \frac{\partial}{\partial \boldsymbol{\theta}} d_\phi(\boldsymbol{\theta}) \Big|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}} + \frac{1}{2} (\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})^T \frac{\partial^2}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T} d_\phi(\boldsymbol{\theta}) \Big|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}} (\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}) + o\left(\|\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}\|^2\right), \quad (45)$$

where

$$\begin{aligned} \frac{\partial}{\partial \boldsymbol{\theta}} d_\phi(\boldsymbol{\theta}) \Big|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}} &= \mathbf{0}_{J-1}, \\ \frac{\partial^2}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T} d_\phi(\boldsymbol{\theta}) \Big|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}} &= \phi''(1) \mathcal{I}_F^{(n_1, n_2)}(\hat{\boldsymbol{\theta}}), \end{aligned}$$

and $\mathcal{I}_F^{(n_1, n_2)}(\boldsymbol{\theta})$ was defined at the beginning of Section 4. Let $\bar{\boldsymbol{\theta}}$ be the parameter vector such that $\bar{\mathbf{p}} = \mathbf{p}(\bar{\boldsymbol{\theta}})$, where $\mathbf{p}(\bar{\boldsymbol{\theta}}) = \mathbf{1}_{2J} \bar{u} + \mathbf{W} \bar{\boldsymbol{\theta}}$, with $\bar{u} = -\log(\mathbf{1}_{2J}^T \exp\{\mathbf{W} \bar{\boldsymbol{\theta}}\})$, is the saturated log-linear model. In particular, for $\boldsymbol{\theta} = \bar{\boldsymbol{\theta}}$ we have

$$d_\phi(\mathbf{p}(\bar{\boldsymbol{\theta}}), \mathbf{p}(\hat{\boldsymbol{\theta}})) = \frac{\phi''(1)}{2} (\bar{\boldsymbol{\theta}} - \hat{\boldsymbol{\theta}})^T \mathcal{I}_F^{(n_1, n_2)}(\hat{\boldsymbol{\theta}}) (\bar{\boldsymbol{\theta}} - \hat{\boldsymbol{\theta}}) + o\left(\|\bar{\boldsymbol{\theta}} - \hat{\boldsymbol{\theta}}\|^2\right).$$

In a similar way it is obtained

$$d_\phi(\mathbf{p}(\bar{\boldsymbol{\theta}}), \mathbf{p}(\hat{\boldsymbol{\theta}}(S))) = \frac{\phi''(1)}{2} (\bar{\boldsymbol{\theta}} - \hat{\boldsymbol{\theta}}(S))^T \mathcal{I}_F^{(n_1, n_2)}(\hat{\boldsymbol{\theta}}(S)) (\bar{\boldsymbol{\theta}} - \hat{\boldsymbol{\theta}}(S)) + o\left(\|\bar{\boldsymbol{\theta}} - \hat{\boldsymbol{\theta}}(S)\|^2\right).$$

Multiplying both sides of the equality by $\frac{2n}{\phi''(1)}$ and taking the difference in both sides of the equality

$$\begin{aligned} T_\phi(\bar{\mathbf{p}}, \mathbf{p}(\hat{\boldsymbol{\theta}}(S)), \mathbf{p}(\hat{\boldsymbol{\theta}})) &= \frac{2n}{\phi''(1)} \left(d_\phi(\mathbf{p}(\bar{\boldsymbol{\theta}}), \mathbf{p}(\hat{\boldsymbol{\theta}})) - d_\phi(\mathbf{p}(\bar{\boldsymbol{\theta}}), \mathbf{p}(\hat{\boldsymbol{\theta}}(S))) \right) \\ &= \sqrt{n} (\bar{\boldsymbol{\theta}} - \hat{\boldsymbol{\theta}})^T \mathcal{I}_F^{(n_1, n_2)}(\hat{\boldsymbol{\theta}}) \sqrt{n} (\bar{\boldsymbol{\theta}} - \hat{\boldsymbol{\theta}}) + o\left(\left\| \sqrt{n} (\bar{\boldsymbol{\theta}} - \hat{\boldsymbol{\theta}}) \right\|^2\right) \\ &\quad - \sqrt{n} (\bar{\boldsymbol{\theta}} - \hat{\boldsymbol{\theta}}(S))^T \mathcal{I}_F^{(n_1, n_2)}(\hat{\boldsymbol{\theta}}(S)) \sqrt{n} (\bar{\boldsymbol{\theta}} - \hat{\boldsymbol{\theta}}(S)) + o\left(\left\| \sqrt{n} (\bar{\boldsymbol{\theta}} - \hat{\boldsymbol{\theta}}(S)) \right\|^2\right). \end{aligned}$$

Now we are going to generalize the three types of estimators by $\hat{\boldsymbol{\theta}}(\bullet)$, understanding that for $\bullet = \emptyset$, $\hat{\boldsymbol{\theta}}(\emptyset) = \bar{\boldsymbol{\theta}}$, $\mathbf{R}(\emptyset) = \mathbf{0}_{(J-1) \times (2J-1)}$, for $\bullet = E$, $\hat{\boldsymbol{\theta}}(E) = \hat{\boldsymbol{\theta}}$, $\mathbf{R}(E) = \mathbf{R}$, and $\bullet = S$, $\hat{\boldsymbol{\theta}}(S)$ and $\mathbf{R}(S)$ as originally defined. It is well-known that

$$\sqrt{n}(\hat{\boldsymbol{\theta}}(\bullet) - \boldsymbol{\theta}_0) = \boldsymbol{\Gamma}(\boldsymbol{\theta}_0, \bullet) \frac{1}{\sqrt{n}} \frac{\partial}{\partial \boldsymbol{\theta}} \ell(\mathbf{N}; \boldsymbol{\theta}) \Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0} + o_p(\mathbf{1}_k), \quad (46)$$

where $\boldsymbol{\theta}_0$ is the true and unknown value of the parameter,

$$\boldsymbol{\Gamma}(\boldsymbol{\theta}_0, \bullet) = \mathcal{I}_F^{-1}(\boldsymbol{\theta}_0) - \mathcal{I}_F^{-1}(\boldsymbol{\theta}_0) \mathbf{R}^T(\bullet) \left(\mathbf{R}(\bullet) \mathcal{I}_F^{-1}(\boldsymbol{\theta}_0) \mathbf{R}^T(\bullet) \right)^{-1} \mathbf{R}(\bullet) \mathcal{I}_F^{-1}(\boldsymbol{\theta}_0), \quad (47)$$

is the variance covariance matrix of $\hat{\boldsymbol{\theta}}(\bullet)$, and $\frac{1}{\sqrt{n}} \frac{\partial}{\partial \boldsymbol{\theta}} \ell(\mathbf{N}; \boldsymbol{\theta}) \Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0} \xrightarrow[n_1, n_2 \rightarrow \infty]{\mathcal{L}} \mathcal{N}(\mathbf{0}_k, \mathcal{I}_F(\boldsymbol{\theta}_0))$ by the Central Limit Theorem. We shall denote

$$\boldsymbol{\Gamma}(\boldsymbol{\theta}_0) = \boldsymbol{\Gamma}(\boldsymbol{\theta}_0, E) = \mathcal{I}_F^{-1}(\boldsymbol{\theta}_0) - \mathcal{I}_F^{-1}(\boldsymbol{\theta}_0) \mathbf{R}^T \left(\mathbf{R} \mathcal{I}_F^{-1}(\boldsymbol{\theta}_0) \mathbf{R}^T \right)^{-1} \mathbf{R} \mathcal{I}_F^{-1}(\boldsymbol{\theta}_0).$$

Taking the differences of both sides of the equality in (46) with cases $\bullet = \emptyset$ and $\bullet = E$, we obtain

$$\sqrt{n}(\bar{\boldsymbol{\theta}} - \hat{\boldsymbol{\theta}}) = (\mathcal{I}_F^{-1}(\boldsymbol{\theta}_0) - \boldsymbol{\Gamma}(\boldsymbol{\theta}_0)) \frac{1}{\sqrt{n}} \frac{\partial}{\partial \boldsymbol{\theta}} \ell(\mathbf{N}; \boldsymbol{\theta}) \Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0} + o_p(\mathbf{1}_k), \quad (48)$$

with cases $\bullet = \emptyset$ and $\bullet = S$,

$$\sqrt{n}(\bar{\boldsymbol{\theta}} - \hat{\boldsymbol{\theta}}(S)) = (\mathcal{I}_F^{-1}(\boldsymbol{\theta}_0) - \boldsymbol{\Gamma}(\boldsymbol{\theta}_0, S)) \frac{1}{\sqrt{n}} \frac{\partial}{\partial \boldsymbol{\theta}} \ell(\mathbf{N}; \boldsymbol{\theta}) \Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0} + o_p(\mathbf{1}_k), \quad (49)$$

and taking into account $\mathcal{I}_F(\hat{\boldsymbol{\theta}}) \xrightarrow[n_1, n_2 \rightarrow \infty]{P} \mathcal{I}_F(\boldsymbol{\theta}_0)$,

$$\begin{aligned} & T_\phi(\bar{\mathbf{p}}, \mathbf{p}(\hat{\boldsymbol{\theta}}(S)), \mathbf{p}(\hat{\boldsymbol{\theta}})) \\ &= \frac{1}{\sqrt{n}} \frac{\partial}{\partial \boldsymbol{\theta}^T} \ell(\mathbf{N}; \boldsymbol{\theta}) \Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0} (\boldsymbol{\Gamma}(\boldsymbol{\theta}_0, S) - \boldsymbol{\Gamma}(\boldsymbol{\theta}_0))^T \mathcal{I}_F(\boldsymbol{\theta}_0) (\boldsymbol{\Gamma}(\boldsymbol{\theta}_0, S) - \boldsymbol{\Gamma}(\boldsymbol{\theta}_0)) \frac{1}{\sqrt{n}} \frac{\partial}{\partial \boldsymbol{\theta}} \ell(\mathbf{N}; \boldsymbol{\theta}) \Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0} + o_p(1) \\ &= \mathbf{Y}^T \mathbf{Y} + o_p(1), \end{aligned} \quad (50)$$

where

$$\mathbf{Y} = \mathbf{A}(\boldsymbol{\theta}_0) (\boldsymbol{\Gamma}(\boldsymbol{\theta}_0, S) - \boldsymbol{\Gamma}(\boldsymbol{\theta}_0)) \mathbf{A}(\boldsymbol{\theta}_0)^T \mathbf{Z},$$

with $\mathbf{Z} \sim \mathcal{N}(\mathbf{0}_{J-1}, \mathbf{I}_{J-1})$ and $\mathbf{A}(\boldsymbol{\theta}_0)$ is the Cholesky's factorization matrix for a non singular matrix such a Fisher information matrix, that is $\mathcal{I}_F(\boldsymbol{\theta}_0) = \mathbf{A}(\boldsymbol{\theta}_0)^T \mathbf{A}(\boldsymbol{\theta}_0)$. In other words

$$\mathbf{Y} \sim \mathcal{N}(\mathbf{0}_k, \mathbf{A}(\boldsymbol{\theta}_0) (\boldsymbol{\Gamma}(\boldsymbol{\theta}_0, S) - \boldsymbol{\Gamma}(\boldsymbol{\theta}_0)) \mathbf{A}(\boldsymbol{\theta}_0)^T),$$

where the variance covariance matrix is idempotent and symmetric. Following Lemma 3 in Ferguson (1996, page 57), $\mathbf{A}(\boldsymbol{\theta}_0) (\boldsymbol{\Gamma}(\boldsymbol{\theta}_0, S) - \boldsymbol{\Gamma}(\boldsymbol{\theta}_0)) \mathbf{A}(\boldsymbol{\theta}_0)^T$ is idempotent and symmetric, if only if $T_\phi(\bar{\mathbf{p}}, \mathbf{p}(\hat{\boldsymbol{\theta}}(S)), \mathbf{p}(\hat{\boldsymbol{\theta}}))$ is a chi-square random variable with degrees of freedom

$$df = \text{rank}(\mathbf{A}(\boldsymbol{\theta}_0) (\boldsymbol{\Gamma}(\boldsymbol{\theta}_0, S) - \boldsymbol{\Gamma}(\boldsymbol{\theta}_0)) \mathbf{A}(\boldsymbol{\theta}_0)^T) = \text{trace}(\mathbf{A}(\boldsymbol{\theta}_0) (\boldsymbol{\Gamma}(\boldsymbol{\theta}_0, S) - \boldsymbol{\Gamma}(\boldsymbol{\theta}_0)) \mathbf{A}(\boldsymbol{\theta}_0)^T).$$

Since

$$(\boldsymbol{\Gamma}(\boldsymbol{\theta}_0, S) - \boldsymbol{\Gamma}(\boldsymbol{\theta}_0))^T \mathcal{I}_F(\boldsymbol{\theta}_0) (\boldsymbol{\Gamma}(\boldsymbol{\theta}_0, S) - \boldsymbol{\Gamma}(\boldsymbol{\theta}_0)) = \boldsymbol{\Gamma}(\boldsymbol{\theta}_0, S) - \boldsymbol{\Gamma}(\boldsymbol{\theta}_0),$$

the condition is reached. The effective degrees of freedom are given by

$$\begin{aligned} df &= \text{trace}(\boldsymbol{\Gamma}(\boldsymbol{\theta}_0, S) \mathbf{A}(\boldsymbol{\theta}_0)^T \mathbf{A}(\boldsymbol{\theta}_0)) - \text{trace}(\boldsymbol{\Gamma}(\boldsymbol{\theta}_0) \mathbf{A}(\boldsymbol{\theta}_0)^T \mathbf{A}(\boldsymbol{\theta}_0)) = \text{trace}(\boldsymbol{\Gamma}(\boldsymbol{\theta}_0, S) \mathcal{I}_F(\boldsymbol{\theta}_0)) - \text{trace}(\boldsymbol{\Gamma}(\boldsymbol{\theta}_0) \mathcal{I}_F(\boldsymbol{\theta}_0)) \\ &= \text{trace}(-\left(\mathbf{R}(S) \mathcal{I}_F^{-1}(\boldsymbol{\theta}_0) \mathbf{R}^T(S)\right)^{-1} \mathbf{R}(S) \mathcal{I}_F^{-1}(\boldsymbol{\theta}_0) \mathbf{R}^T(S)) \\ &\quad - \text{trace}(-\left(\mathbf{R} \mathcal{I}_F^{-1}(\boldsymbol{\theta}_0) \mathbf{R}^T\right)^{-1} \mathbf{R} \mathcal{I}_F^{-1}(\boldsymbol{\theta}_0) \mathbf{R}^T) \\ &= (J - 1) - \text{card}(S). \end{aligned}$$

Regarding the other test-statistic $S_\phi(\mathbf{p}(\hat{\boldsymbol{\theta}}(S)), \mathbf{p}(\hat{\boldsymbol{\theta}}))$, observe that if we take (45), in particular for $\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}(S)$ it is obtained

$$d_\phi(\hat{\boldsymbol{\theta}}(S)) = \frac{\phi''(1)}{2} (\hat{\boldsymbol{\theta}}(S) - \hat{\boldsymbol{\theta}})^T \mathcal{I}_F(\hat{\boldsymbol{\theta}}) (\hat{\boldsymbol{\theta}}(S) - \hat{\boldsymbol{\theta}}) + o\left(\left\|\hat{\boldsymbol{\theta}}(S) - \hat{\boldsymbol{\theta}}\right\|^2\right).$$

In addition, (48)–(49) is

$$\sqrt{n}(\hat{\boldsymbol{\theta}}(S) - \hat{\boldsymbol{\theta}}) = (\boldsymbol{\Gamma}(\boldsymbol{\theta}_0, S) - \boldsymbol{\Gamma}(\boldsymbol{\theta}_0)) \frac{1}{\sqrt{n}} \frac{\partial}{\partial \boldsymbol{\theta}} \ell(\mathbf{N}; \boldsymbol{\theta}) \Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0} + o_p(\mathbf{1}_k),$$

and taking into account $\mathcal{I}_F(\hat{\boldsymbol{\theta}}) \xrightarrow[n_1, n_2 \rightarrow \infty]{P} \mathcal{I}_F(\boldsymbol{\theta}_0)$ and (50), it follows (44), which means from Slutsky's Theorem that both test-statistics have the same asymptotic distribution. ■

A.2 Lemma

Let \mathbf{Y} be a k -dimensional random variable with normal distribution $\mathcal{N}(\mathbf{0}_k, \mathbf{Q})$ with \mathbf{Q} being a projection matrix, that is idempotent and symmetric, and let \mathbf{d}_i be the fixed k -dimensional vectors such that for them either $\mathbf{Q}\mathbf{d}_i = \mathbf{0}_k$ or $\mathbf{Q}\mathbf{d}_i = \mathbf{d}_i$, $i = 1, \dots, k$, is true. Then $\left(\mathbf{Y}^T \mathbf{Y} \mid \mathbf{d}_i^T \mathbf{Y} \geq 0, i = 1, \dots, k\right) \sim \chi_{df}^2$, where $df = \text{rank}(\mathbf{Q})$.

Proof. This result can be found in several sources, for instance in Kudô (1963, page 414), Barlow et al. (1972, page 128) and Shapiro (1985, page 139). ■

A.3 Proof of Theorem 2

We shall perform the proof for $S_\phi(\mathbf{p}(\tilde{\boldsymbol{\theta}}), \mathbf{p}(\hat{\boldsymbol{\theta}}))$. It suppose that it is true $\mathbf{R}\boldsymbol{\theta} \geq \mathbf{0}_{J-1}$ and we want to test $\mathbf{R}\boldsymbol{\theta} = \mathbf{0}_{J-1}$ (H_0). It is clear that if H_0 is not true is because there exists some index $i \in E$ such that $\mathbf{R}(\{i\})\boldsymbol{\theta} > 0$. Let us consider the family of all possible subsets in E , denoted by $\mathcal{F}(E)$, then we shall specify more thoroughly $\tilde{\boldsymbol{\theta}}$ by $\tilde{\boldsymbol{\theta}}(S)$ when there exists $S \in \mathcal{F}(E)$ such that

$$\mathbf{R}(S)\tilde{\boldsymbol{\theta}} = \mathbf{0}_{\text{card}(S)} \quad \text{and} \quad \mathbf{R}(S^C)\tilde{\boldsymbol{\theta}} > \mathbf{0}_{(J-1)-\text{card}(S)}.$$

It is clear that for a sample $\tilde{\boldsymbol{\theta}} = \tilde{\boldsymbol{\theta}}(S)$ can be true only for a unique set of indices $S \in \mathcal{F}(E)$, and thus by applying the Theorem of Total Probability

$$\Pr\left(S_\phi(\mathbf{p}(\tilde{\boldsymbol{\theta}}), \mathbf{p}(\hat{\boldsymbol{\theta}})) \leq x\right) = \sum_{S \in \mathcal{F}(E)} \Pr\left(S_\phi(\mathbf{p}(\tilde{\boldsymbol{\theta}}), \mathbf{p}(\hat{\boldsymbol{\theta}})) \leq x, \tilde{\boldsymbol{\theta}} = \tilde{\boldsymbol{\theta}}(S)\right).$$

From the Karush-Khun-Tucker necessary conditions (see for instance Theorem 4.2.13 in Bazaraa et al. (2006)) to solve the optimization problem $\max \ell(\mathbf{N}; \boldsymbol{\theta})$ s.t. $\mathbf{R}\boldsymbol{\theta} \geq \mathbf{0}_{J-1}$, associated with $\tilde{\boldsymbol{\theta}}$,

$$\frac{\partial}{\partial \boldsymbol{\theta}} \ell(\mathbf{N}; \boldsymbol{\theta}) + \sum_{i=1}^{J-1} \lambda_i \mathbf{R}^T(\{i\}) = 0, \quad i = 1, \dots, J-1, \quad (51a)$$

$$\lambda_i \mathbf{R}(\{i\})\boldsymbol{\theta} = 0, \quad i = 1, \dots, J-1, \quad (51b)$$

$$\lambda_i \leq 0, \quad i = 1, \dots, J-1, \quad (51c)$$

the only conditions which characterize the MLE $\tilde{\boldsymbol{\theta}} = \tilde{\boldsymbol{\theta}}(S)$ with a specific $S \in \mathcal{F}(E)$, are the complementary slackness conditions $\mathbf{R}(\{i\})\boldsymbol{\theta} > 0$, for $i \in S$ and $\lambda_i < 0$, for $i \in S^C$, since $\frac{\partial}{\partial \boldsymbol{\theta}} \ell(\mathbf{N}; \boldsymbol{\theta}) + \lambda_i \mathbf{R}^T(\{i\}) = 0$, $i = 1, \dots, J-1$, $\mathbf{R}(\{i\})\boldsymbol{\theta} = 0$, for $i \in S^C$ and $\lambda_i = 0$, for $i \in S$ are redundant conditions once we know that the Karush-Khun-Tucker necessary conditions are true for all the possible sets $S \in \mathcal{F}(E)$ which define $\tilde{\boldsymbol{\theta}} = \tilde{\boldsymbol{\theta}}(S)$. For this reason we can consider

$$\begin{aligned} & \Pr\left(S_\phi(\mathbf{p}(\tilde{\boldsymbol{\theta}}), \mathbf{p}(\hat{\boldsymbol{\theta}})) \leq x, \tilde{\boldsymbol{\theta}} = \tilde{\boldsymbol{\theta}}(S)\right) = \\ & \Pr\left(S_\phi(\mathbf{p}(\tilde{\boldsymbol{\theta}}), \mathbf{p}(\hat{\boldsymbol{\theta}})) \leq x, \tilde{\boldsymbol{\lambda}}(S) < \mathbf{0}_{\text{card}(S)}, \mathbf{R}(S^C)\tilde{\boldsymbol{\theta}}(S) > \mathbf{0}_{(J-1)-\text{card}(S)}\right), \end{aligned}$$

where $\tilde{\boldsymbol{\lambda}}(S)$ is the vector of the vector of Karush-Khun-Tucker multipliers associated with estimator $\tilde{\boldsymbol{\theta}}(S)$. Furthermore, under H_0 , $\mathbf{R}\tilde{\boldsymbol{\theta}}(S) = \mathbf{R}\tilde{\boldsymbol{\theta}}(S) - \mathbf{R}\boldsymbol{\theta}_0$, because $\mathbf{R}\boldsymbol{\theta}_0 = \mathbf{0}_{J-1}$, hence

$$\Pr\left(S_\phi(\mathbf{p}(\tilde{\boldsymbol{\theta}}), \mathbf{p}(\hat{\boldsymbol{\theta}})) \leq x\right) = \sum_{S \in \mathcal{F}(E)} \Pr\left(S_\phi(\mathbf{p}(\tilde{\boldsymbol{\theta}}), \mathbf{p}(\hat{\boldsymbol{\theta}})) \leq x, \tilde{\boldsymbol{\lambda}}(S) < \mathbf{0}_{\text{card}(S)}, \mathbf{R}(S^C)\tilde{\boldsymbol{\theta}}(S) - \mathbf{R}(S^C)\boldsymbol{\theta}_0 > \mathbf{0}_{\text{card}(S^C)}\right),$$

where $\text{card}(S^C) = (J-1) - \text{card}(S)$. On the other hand, (51a) and (51b) are also true for $(\hat{\boldsymbol{\theta}}^T(S), \hat{\boldsymbol{\lambda}}^T(S))^T$ according to the Lagrange multipliers method. Hence, $\tilde{\boldsymbol{\theta}}(S) = \hat{\boldsymbol{\theta}}(S)$ and $\tilde{\boldsymbol{\lambda}}(S) = \hat{\boldsymbol{\lambda}}(S)$. It follows that:

- under $\tilde{\boldsymbol{\theta}} = \hat{\boldsymbol{\theta}}(S)$, $S_\phi(\mathbf{p}(\tilde{\boldsymbol{\theta}}), \mathbf{p}(\hat{\boldsymbol{\theta}})) = S_\phi(\mathbf{p}(\hat{\boldsymbol{\theta}}(S)), \mathbf{p}(\hat{\boldsymbol{\theta}}))$ and taking into account Proposition A.1

$$\begin{aligned} S_\phi(\mathbf{p}(\tilde{\boldsymbol{\theta}}), \mathbf{p}(\hat{\boldsymbol{\theta}})) &= T_\phi(\bar{\mathbf{p}}, \mathbf{p}(\hat{\boldsymbol{\theta}}(S)), \mathbf{p}(\hat{\boldsymbol{\theta}})) + o_p(1) \\ &= \left(\mathbf{A}(\boldsymbol{\theta}_0) (\boldsymbol{\Gamma}(\boldsymbol{\theta}_0, S) - \boldsymbol{\Gamma}(\boldsymbol{\theta}_0)) \mathbf{A}(\boldsymbol{\theta}_0)^T \mathbf{Z} \right)^T \left(\mathbf{A}(\boldsymbol{\theta}_0) (\boldsymbol{\Gamma}(\boldsymbol{\theta}_0, S) - \boldsymbol{\Gamma}(\boldsymbol{\theta}_0)) \mathbf{A}(\boldsymbol{\theta}_0)^T \mathbf{Z} \right) + o_p(1), \\ &= \mathbf{Z}^T \mathbf{A}(\boldsymbol{\theta}_0) (\boldsymbol{\Gamma}(\boldsymbol{\theta}_0, S) - \boldsymbol{\Gamma}(\boldsymbol{\theta}_0)) \mathbf{A}(\boldsymbol{\theta}_0)^T \mathbf{Z} + o_p(1). \end{aligned}$$

where $\mathbf{Z} \sim \mathcal{N}(\mathbf{0}_k, \mathbf{I}_k)$.

- under $\tilde{\boldsymbol{\lambda}}(S) = \hat{\boldsymbol{\lambda}}(S)$ and from Sen et al. (2010, page 267 formula (8.6.28))

$$\begin{aligned} \frac{1}{\sqrt{n}} \tilde{\boldsymbol{\lambda}}(S) &= \sqrt{n} \mathbf{Q}^T(\boldsymbol{\theta}_0, S) \frac{1}{\sqrt{n}} \frac{\partial}{\partial \boldsymbol{\theta}} \ell(\mathbf{N}; \boldsymbol{\theta}) \Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0} + o_p(\mathbf{1}_{\text{card}(S)}) \\ &= \mathbf{Q}^T(\boldsymbol{\theta}_0, S) \mathbf{A}(\boldsymbol{\theta}_0)^T \mathbf{Z} + o_p(\mathbf{1}_{\text{card}(S)}), \end{aligned}$$

where

$$\mathbf{Q}(\boldsymbol{\theta}_0, S) = -\mathcal{I}_F^{-1}(\boldsymbol{\theta}_0) \mathbf{R}^T(S) \mathbf{L}(\boldsymbol{\theta}_0, S) \left(\mathbf{R}(S) \mathcal{I}_F^{-1}(\boldsymbol{\theta}_0) \mathbf{R}^T(S) \right)^{-1};$$

- under $\tilde{\boldsymbol{\theta}} = \hat{\boldsymbol{\theta}}(S)$ and from (46)

$$\begin{aligned} \sqrt{n} \left(\mathbf{R}(S^C) \tilde{\boldsymbol{\theta}}(S) - \mathbf{R}(S^C) \boldsymbol{\theta}_0 \right) &= \sqrt{n} \mathbf{R}(S^C) \boldsymbol{\Gamma}(\boldsymbol{\theta}_0, S) \frac{1}{\sqrt{n}} \frac{\partial}{\partial \boldsymbol{\theta}} \ell(\mathbf{N}; \boldsymbol{\theta}) \Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0} + o_p(\mathbf{1}_{\text{card}(S^C)}) \\ &= \mathbf{R}(S^C) \boldsymbol{\Gamma}(\boldsymbol{\theta}_0, S) \mathbf{A}(\boldsymbol{\theta}_0)^T \mathbf{Z} + o_p(\mathbf{1}_{\text{card}(S^C)}). \end{aligned}$$

That is,

$$\begin{aligned} \lim_{n_1, n_2 \rightarrow \infty} \Pr \left(S_\phi(\mathbf{p}(\tilde{\boldsymbol{\theta}}), \mathbf{p}(\hat{\boldsymbol{\theta}})) \leq x \right) &= \sum_{S \in \mathcal{F}(E)} \Pr \left(\mathbf{Z}_3^T(S) \mathbf{Z}_3(S) \leq x, \mathbf{Z}_1(S) \geq \mathbf{0}_{\text{card}(S)}, \mathbf{Z}_2(S) \geq \mathbf{0}_{\text{card}(S^C)} \right) \\ &= \sum_{S \in \mathcal{F}(E)} \Pr \left(\mathbf{Z}_3^T(S) \mathbf{Z}_3(S) \leq x \mid \mathbf{Z}_1(S) \geq \mathbf{0}_{\text{card}(S)}, \mathbf{Z}_2(S) \geq \mathbf{0}_{\text{card}(S^C)} \right) \Pr \left(\mathbf{Z}_1(S) \geq \mathbf{0}_{\text{card}(S)}, \mathbf{Z}_2(S) \geq \mathbf{0}_{\text{card}(S^C)} \right) \\ &= \sum_{S \in \mathcal{F}(E)} \Pr \left(\mathbf{Z}_3^T(S) \mathbf{Z}_3(S) \leq x \mid \left(\mathbf{Z}_1^T(S), \mathbf{Z}_2^T(S) \right)^T \geq \mathbf{0}_{J-1} \right) \Pr \left(\mathbf{Z}_1(S) \geq \mathbf{0}_{\text{card}(S)}, \mathbf{Z}_2(S) \geq \mathbf{0}_{\text{card}(S^C)} \right), \end{aligned}$$

where

$$\begin{aligned} \mathbf{Z}_3(S) &= \mathbf{M}_3(\boldsymbol{\theta}_0, S) \mathbf{Z}, & \mathbf{M}_3(\boldsymbol{\theta}_0, S) &= \mathbf{A}(\boldsymbol{\theta}_0) (\boldsymbol{\Gamma}(\boldsymbol{\theta}_0, S) - \boldsymbol{\Gamma}(\boldsymbol{\theta}_0)) \mathbf{A}(\boldsymbol{\theta}_0)^T, \\ \mathbf{Z}_1(S) &= \mathbf{M}_1(\boldsymbol{\theta}_0, S) \mathbf{Z}, & \mathbf{M}_1(\boldsymbol{\theta}_0, S) &= -\mathbf{Q}^T(\boldsymbol{\theta}_0, S) \mathbf{A}(\boldsymbol{\theta}_0)^T, \\ \mathbf{Z}_2(S) &= \mathbf{M}_2(\boldsymbol{\theta}_0, S) \mathbf{Z}, & \mathbf{M}_2(\boldsymbol{\theta}_0, S) &= \mathbf{R}(S^C) \boldsymbol{\Gamma}(\boldsymbol{\theta}_0, S) \mathbf{A}(\boldsymbol{\theta}_0)^T. \end{aligned}$$

Taking into account that $\mathbf{M}_3(\boldsymbol{\theta}_0, S) \mathbf{M}_2^T(\boldsymbol{\theta}_0, S) = \mathbf{M}_2^T(\boldsymbol{\theta}_0, S)$ and $\mathbf{M}_3(\boldsymbol{\theta}_0, S) \mathbf{M}_1^T(\boldsymbol{\theta}_0, S) = \mathbf{0}_{(J-1) \times \text{card}(S)}$, by applying the lemma given in Section A.2

$$\Pr \left(\mathbf{Z}_3^T(S) \mathbf{Z}_3(S) \leq x \mid \left(\mathbf{Z}_1^T(S), \mathbf{Z}_2^T(S) \right)^T \geq \mathbf{0}_{J-1} \right) = \Pr \left(\chi_{df}^2 \leq x \right)$$

where

$$\begin{aligned} df &= \text{rank} \left(\mathbf{A}(\boldsymbol{\theta}_0) (\boldsymbol{\Gamma}(\boldsymbol{\theta}_0, S) - \boldsymbol{\Gamma}(\boldsymbol{\theta}_0)) \mathbf{A}(\boldsymbol{\theta}_0)^T \right) = \text{trace} \left(\mathbf{A}(\boldsymbol{\theta}_0) (\boldsymbol{\Gamma}(\boldsymbol{\theta}_0, S) - \boldsymbol{\Gamma}(\boldsymbol{\theta}_0)) \mathbf{A}(\boldsymbol{\theta}_0)^T \right) \\ &= (J-1) - \text{card}(S). \end{aligned}$$

Finally,

$$\begin{aligned}
& \lim_{n_1, n_2 \rightarrow \infty} \Pr \left(S_\phi(\mathbf{p}(\tilde{\boldsymbol{\theta}}), \mathbf{p}(\hat{\boldsymbol{\theta}})) \leq x \right) \\
&= \sum_{S \in \mathcal{F}(E)} \Pr \left(\chi_{(J-1)-\text{card}(S)}^2 \leq x \right) \Pr \left(\mathbf{Z}_1(S) \geq \mathbf{0}_{\text{card}(S)}, \mathbf{Z}_2(S) \geq \mathbf{0}_{\text{card}(S^C)} \right) \\
&= \sum_{j=0}^{J-1} \Pr \left(\chi_{(J-1)-j}^2 \leq x \right) \sum_{S \in \mathcal{F}(E), \text{card}(S)=j} \Pr \left(\mathbf{Z}_1(S) \geq \mathbf{0}_{\text{card}(S)}, \mathbf{Z}_2(S) \geq \mathbf{0}_{\text{card}(S^C)} \right),
\end{aligned}$$

and since $\mathbf{Q}^T(\boldsymbol{\theta}_0, S) \mathcal{I}_F(\boldsymbol{\theta}_0) \boldsymbol{\Gamma}(\boldsymbol{\theta}_0, S) = \mathbf{0}_{\text{card}(S) \times (J-1)}$, it holds $\mathbf{M}_1(\boldsymbol{\theta}_0, S) \mathbf{M}_2^T(\boldsymbol{\theta}_0, S) = \mathbf{0}_{\text{card}(S) \times \text{card}(S^C)}$ which means that $\mathbf{Z}_1(S)$ and $\mathbf{Z}_2(S)$ are independent, that is

$$\lim_{n_1, n_2 \rightarrow \infty} \Pr \left(S_\phi(\mathbf{p}(\tilde{\boldsymbol{\theta}}), \mathbf{p}(\hat{\boldsymbol{\theta}})) \leq x \right) = \sum_{j=0}^{J-1} \Pr \left(\chi_{(J-1)-j}^2 \leq x \right) w_j(\boldsymbol{\theta}_0)$$

where the expression of $w_j(\boldsymbol{\theta}_0)$ is (30). We have also,

$$\text{Var}(\mathbf{Z}_1(S)) = \mathbf{M}_1(\boldsymbol{\theta}_0, S) \mathbf{M}_1^T(\boldsymbol{\theta}_0, S) = \mathbf{Q}^T(\boldsymbol{\theta}_0, S) \mathcal{I}_F(\boldsymbol{\theta}_0) \mathbf{Q}(\boldsymbol{\theta}_0, S) = \left(\mathbf{R}(S) \mathcal{I}_F^{-1}(\boldsymbol{\theta}_0) \mathbf{R}^T(S) \right)^{-1} = \mathbf{H}^{-1}(S, S, \boldsymbol{\theta}_0),$$

$$\begin{aligned}
\text{Var}(\mathbf{Z}_2(S)) &= \mathbf{M}_2(\boldsymbol{\theta}_0, S) \mathbf{M}_2^T(\boldsymbol{\theta}_0, S) = \mathbf{R}(S^C) \boldsymbol{\Gamma}(\boldsymbol{\theta}_0, S) \mathcal{I}_F(\boldsymbol{\theta}_0) \boldsymbol{\Gamma}^T(\boldsymbol{\theta}_0, S) \mathbf{R}^T(S^C) = \mathbf{R}(S^C) \boldsymbol{\Gamma}(\boldsymbol{\theta}_0, S) \mathbf{R}^T(S^C) \\
&= \mathbf{H}(S^C, S^C, \boldsymbol{\theta}_0) - \mathbf{H}(S^C, S, \boldsymbol{\theta}_0) \mathbf{H}^{-1}(S, S, \boldsymbol{\theta}_0) \mathbf{H}^T(S^C, S, \boldsymbol{\theta}_0).
\end{aligned}$$

The proof of $T_\phi(\bar{\mathbf{p}}, \mathbf{p}(\tilde{\boldsymbol{\theta}}), \mathbf{p}(\hat{\boldsymbol{\theta}}))$ is almost immediate from the proof for $S_\phi(\mathbf{p}(\tilde{\boldsymbol{\theta}}), \mathbf{p}(\hat{\boldsymbol{\theta}}))$ and taking into account that for some $S \in \mathcal{F}(E)$

$$T_\phi(\bar{\mathbf{p}}, \mathbf{p}(\tilde{\boldsymbol{\theta}}), \mathbf{p}(\hat{\boldsymbol{\theta}})) = T_\phi(\bar{\mathbf{p}}, \mathbf{p}(\hat{\boldsymbol{\theta}}(S)), \mathbf{p}(\hat{\boldsymbol{\theta}})) + o_p(1) = S_\phi(\mathbf{p}(\tilde{\boldsymbol{\theta}}), \mathbf{p}(\hat{\boldsymbol{\theta}})).$$